Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting EGFR-TKI resistance in EGFR-mutant lung adenocarcinoma from H&E histopathology with weakly supervised multiple-instance learning

Weiwei Wang
Hebei Medical University
Yan Zhang
wang03281994@163.com

Hebei Medical University
Peng Zhao
Hebei Medical University

Research Article

Keywords: Lung adenocarcinoma, EGFR-TKI resistance, Computational pathology, Deep learning, Multiple instance learning

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10300980/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting EGFR-TKI resistance in EGFR-mutant lung adenocarcinoma from H&E histopathology with weakly supervised multiple-instance learning

Weiwei Wang[1,3], Peng Zhao[2,4], Zhang Yan[1,5*]

[1] Department of Oncology, Graduate School, Hebei Medical University , Shijiazhuang, Hebei, China

[2] Department of Neurosurgery, Graduate School, Hebei Medical University ,Shijiazhuang, Hebei, China

[3] Department of Pathology, Hengshui People's Hospital, Hengshui, Hebei, China

[4] Department of Neurosurgery, Hengshui People's Hospital, Hengshui, Hebei, China

[5] Department of Oncology, Shijiazhuang People's Hospital, Shijiazhuang, Hebei, China

*Correspondence: wang03281994@163.com (Zhang Yan)

Abstract

Background: This study sought to construct a deep-learning pipeline using an attention mechanism and multiple-instance learning (MIL) to predict early treatment response to epidermal growth factor receptor-tyrosine kinase inhibitors (EGFR-TKIs) directly from routinely available hematoxylin-eosin (H&E) whole-slide images (WSIs) of pretreatment tissue in patients with EGFR-mutant lung adenocarcinoma (LUAD).

Methods: We retrospectively obtained clinicopathological data from 100 patients with EGFR-mutant lung adenocarcinoma. The framework employed ResNet-50 for feature representation and integrated a gated-attention pooling mechanism for feature aggregation at the patient level. This weakly supervised approach does not require pixel-level manual annotations and enables direct slide-level prediction of drug resistance.

Results: The proposed gated-attention framework attained an area under the receiver operating characteristic curve (AUC) of 0.891 in 5-fold cross-validation, outperforming conventional mean-pooling, max-pooling methods, and Transformer-based architectures such as TransMIL. Ablation analyses showed that the explicit identification of discriminative histopathological areas is essential for capturing the spatial variability within tumor specimens. Using Grad-CAM visualization and HoVer-Net-based nuclear morphometric analysis, the model captured a latent morphologic risk phenotype associated with drug resistance. The resistant group exhibited higher cellular density, greater architectural disorder, greater nuclear areas (78.3 ± 22.1 μm^2), higher nuclear-to-cytoplasmic ratios, and higher chromatin-texture entropy. In addition, this phenotype was closely associated with TP53 co-mutation enrichment (62.5% in the resistant group vs. 38.3% in the sensitive group), indicating underlying genomic instability.

Conclusion: This study demonstrates that a lightweight and weakly supervised computational-pathology framework can support the early estimation of EGFR-TKI treatment response. This pipeline is noninvasive and scalable, provides biologically explainable outputs, and may offer a practical clinical decision-support tool for precision oncology.

Keywords: Lung adenocarcinoma; EGFR-TKI resistance; Computational pathology; Deep learning; Multiple instance learning

Introduction

Lung cancer continues to be the leading cause of cancer-related mortality worldwide, with lung adenocarcinoma (LUAD) representing its most common histologic subtype [30]. Activating mutations in the epidermal growth factor receptor (EGFR), mainly exon 19 deletions and L858R point mutations, define a clinically actionable molecular subset of LUAD [31, 34]. EGFR tyrosine kinase inhibitors (EGFR-TKIs) have substantially improved outcomes in this population, producing notable gains in progression-free survival and overall survival [1, 2, 17]. However, nearly all cases eventually experience disease progression due to intrinsic or acquired resistance, thereby constraining long-term therapeutic benefit.

Mechanisms of EGFR-TKI resistance are heterogeneous, and include secondary EGFR mutations (e.g., T790M, C797S), activation of bypass pathways (e.g., MET or HER2 amplification), epithelial–mesenchymal transition, and remodeling of the tumor microenvironment [3, 5]. Current clinical practice largely relies on tissue- or liquid-biopsy-based molecular testing assays including next-generation sequencing and PCR to detect resistance-associated mechanisms [6–8]. While clinically useful, these approaches remain

constrained by invasiveness, cost, sampling bias, and limited ability to capture spatiotemporal tumor variability [9].

Although radiologic imaging and radiomics have been explored for EGFR mutation prediction, they remain inadequate for resolving cellular and tissue-level histopathologic variability that underlies treatment response [12, 34]. Therefore, there is an unmet clinical need for a scalable, noninvasive, and cost-effective approach capable of predicting EGFR-TKI treatment response directly from routinely available hematoxylin-eosin (H&E)-stained WSIs [9, 10].

Recent advances in computational pathology and deep learning have enabled extraction of clinically relevant molecular and prognostic signals from H&E images [10, 13, 33]. Specifically, weakly supervised multiple-instance learning (MIL) frameworks have become a leading paradigm for WSI analysis because they do not require pixel-level annotations [13–15, 10]. However, most previous studies have focused on mutation-status estimation rather than clinically actionable outcomes such as treatment response or resistance phenotype [10, 33]. In addition, the biological interpretability of deep-learning-derived histology remains incompletely understood.

Here, we developed an attention-based MIL pipeline for predicting EGFR-TKI resistance directly from pretreatment H&E WSIs of EGFR-mutant LUAD with ResNet-50-based feature representation and gated-attention pooling [19, 13], enabling slide-level estimation without pixel-level annotations. We further employed Grad-CAM visualization and HoVer-Net-based nuclear morphometric profiling to characterize histologic features underlying framework predictions [25, 27]. Overall, we aimed to develop

an explainable computational pathology model for early stratification of EGFR-TKI response and link histomorphologic patterns to tumor biology.

Methods

Patient cohort and tissue specimens

We retrospectively obtained H&E-stained histopathologic specimens from 100 patients with EGFR-mutant lung adenocarcinoma at two centers: Shijiazhuang People0 patienerlying Hengshui Peoplei Peoplent lung adenocarcinoma at two centers: ShAll included patients had undergone surgical resection or biopsy, and EGFR mutation status was verified by PCR-based assays or NGS. Treatment response was evaluated according to RECIST version 1.1 criteria at 12 weeks after EGFR-TKI initiation. Demographic characteristics, smoking history, specimen type, clinical stage, predominant histologic pattern, molecular testing method, EGFR mutation subtype, first-line EGFR-TKI regimen, TP53 co-mutation status, and Ki-67 labeling index were derived from electronic medical records and the pathology information system.

The cohort was categorized into two groups: (1) EGFR-TKI resistant ($n = 40$), defined as cases with EGFR-sensitizing mutations who showed progressive disease or stable disease with a best response of less than 10% reduction during the first 12 weeks of TKI therapy; and (2) EGFR-TKI sensitive ($n = 60$), defined as patients with verified EGFR-sensitizing mutations (exon 19 deletion or L858R) who achieved a partial or complete response. For each patient, formalin-fixed paraffin-embedded (FFPE) tissue blocks were sectioned and stained using a standard H&E protocol.

Digital pathology and image preprocessing

Stained tissue sections were scanned with an Aperio AT2 whole-slide scanner at 40×× magnification (0.25 μμm/pixel) [9, 28]. Quality control was conducted by a board-certified pathologist (X.C.) who excluded slides showing substantial artifacts, tissue folding, or defocused areas.

Tissue areas were separated from background with Otsuue areas were separatedapplied to the saturation channel in HSV color space. Nonoverlapping patches of 512×512 pixels were derived from segmented tissue regions at 20×× equivalent magnification (0.5 μμm/pixel) [29, 9]. Patches with less than 70% tissue content were excluded. Color normalization was performed using the Macenko method to minimize interslide staining variability, with a reference image chosen from slides with median staining intensity.

For each patient, we derived a mean of 345 ± 128 H&E patches. During training, 100 patches were randomly sampled for each patient in each epoch; all accessible patches were used for inference.

Model architecture

The proposed architecture (Fig. 1) uses a patient-level, single-stream H&E workflow. Tissue areas are first segmented from the WSI, extracted as H&E patches, augmented, normalized with the Macenko method, and processed with a ResNet-50 feature encoder. Patch-level embeddings are then aggregated by gated-attention MIL into a slide-level representation, which is subsequently passed through a two-layer MLP, dropout regularization, and a linear classifier to generate a sigmoid resistance probability. The

trained attention scores are preserved as an attention heatmap for framework interpretability.

Feature extraction

The framework uses a ResNet-50 backbone pretrained on ImageNet [28, 16], with the final fully connected layer removed. For each patient, let $\{x_i\}_{i=1}^{N}$ denote the set of H&E patches. Each patch is processed independently by the encoder to generate a 2048-dimensional feature representation:

$$h_i = f(x_i; \theta) \in \mathbb{R}^{2048} \tag{1}$$

Gated attention pooling

Patch-level features are aggregated to form a slide-level representation with gated-attention pooling:

$$a_i = \frac{\exp\{\mathbf{w}^T(\tanh(\mathbf{V}h_i) \odot \mathrm{sigm}(\mathbf{U}h_i))\}}{\sum_{k=1}^{N} \exp\{\mathbf{w}^T(\tanh(\mathbf{V}h_k) \odot \mathrm{sigm}(\mathbf{U}h_k))\}} \tag{2}$$

$$z = \sum_{i=1}^{N} a_i \, h_i \in \mathbb{R}^{2048} \tag{3}$$

where $\mathbf{V} \in \mathbb{R}^{256\times2048}$, $\mathbf{U} \in \mathbb{R}^{256\times2048}$, and $\mathbf{w} \in \mathbb{R}^{256}$ are trainable parameters. The gating mechanism (sigmoid) modulates the attention signal (tanh), allowing the framework to down-weight uninformative patches.

Classification head

The slide-level representation z is processed using a two-layer MLP with a hidden dimension of 256 and ReLU activation:

$$z_{out} = W_2 \cdot \mathrm{ReLU}(W_1 z + b_1) + b_2 \tag{4}$$

where $W_1 \in \mathbb{R}^{256\times2048}$ and $W_2 \in \mathbb{R}^{256\times256}$. A dropout layer ($p = 0.5$) regularizes the MLP output before a linear classifier generates a single scalar logit, which is converted by a sigmoid function into the EGFR-TKI resistance probability:

$$\hat{y} = w_c^T z_{out} + b_c \tag{5}$$

The framework is trained with binary cross-entropy loss with class-balanced weighting to account for class imbalance:

$$\mathcal{L} = -[w_R \cdot y\log(\sigma(\hat{y})) + w_S \cdot (1-y)\log(1-\sigma(\hat{y}))] \tag{6}$$

where σ is the sigmoid function, $w_R = N_{total}/(2 \cdot N_R)$, and $w_S = N_{total}/(2 \cdot N_S)$.

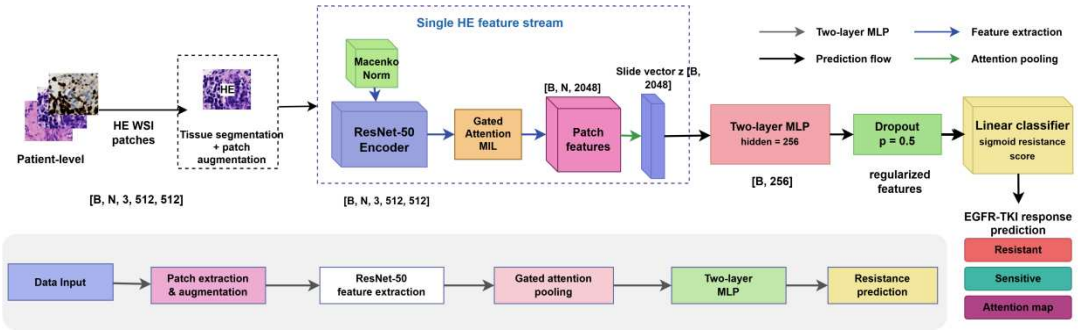


Fig. 1 Overview of the proposed single-stream H&E attention MIL architecture. Patient-level H&E WSIs are segmented into tissue patches, augmented, and Macenko-normalized

before being passed into a ResNet-50 encoder. The encoder generates patch embeddings with shape $[B, N, 2048]$, which are subsequently aggregated by gated-attention MIL into a slide-level vector z with shape $[B, 2048]$. The slide-level vector is passed through a two-layer MLP (a hidden dimension of 256), dropout regularization ($p = 0.5$), and a linear classifier to produce a sigmoid EGFR-TKI resistance probability. The output is interpreted as a resistant or sensitive response, while the learned attention scores provide attention maps for visualizing informative H&E regions.

Training protocol

We used 5-fold stratified cross-validation, preserving the class proportions within each fold. Data augmentation comprised random horizontal and vertical flips, random rotations (0°, 90°, 180°, 270°), color jitter (brightness ± 0.1, contrast ± 0.1, saturation ± 0.05), and random erasing at a probability of 0.1.

The framework was optimized with AdamW with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and a cosine-annealing schedule with warm restarts over 80 epochs [16, 20]. Batch size was set to 16 cases per iteration. Early stopping with a patience period of 15 epochs was applied according to validation AUC [14]. All experiments were performed on a single NVIDIA A100 GPU (40 GB) with PyTorch 2.0 and the timm library.

Baseline models

To assess the importance of the gated-attention pooling mechanism, we compared it with the following MIL aggregation methods, all using the same ResNet-50 backbone:

(a) *Mean pooling*: global average pooling of patch features as the slide-level representation.

(b) *Max pooling*: element-wise maximum across all patch features.

(c) *ABMIL*: standard attention-based MIL without gating [13].

(d) *CLAM*: clustering-constrained attention MIL with instance-level clustering loss [14].

(e) *TransMIL*: Transformer-based MIL using self-attention across patch tokens with positional encoding [15, 20].

We also assessed alternative backbone architectures, including DenseNet-121, EfficientNet-B3, and a Vision Transformer (ViT-B/16) pretrained on ImageNet-21k [16, 20].

Evaluation metrics

Model performance was assessed with the following measures calculated on held-out test folds: AUC, balanced accuracy, F1 score, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Ninety-five percent confidence intervals were estimated with 1,000 bootstrap resamples of the test-set predictions. Pairwise statistical comparisons between models were performed using DeLong test for differences in AUC [24].

Model interpretation

We used gradient-weighted class activation mapping (Grad-CAM) [25] to produce spatial activation heatmaps identifying H&E patch regions that contributed to classification. The

heatmaps and gated-attention scores were projected onto WSI coordinates to display discriminative tissue areas, with attention scores providing a quantitative estimate of each patch's contribution to the slide-level prediction.

To characterize class-associated morphologic features, we applied t-distributed stochastic neighbor embedding (t-SNE) to the trained slide-level representations (z) from all cases in the test folds. Nuclear segmentation was conducted on high-attention H&E patches with HoVer-Net [27] to extract quantitative nuclear morphometric features (nuclear size, eccentricity, chromatin texture, and nuclear-to-cytoplasmic ratio) for correlation with model-based estimations.

Results

Patient characteristics

The clinical, pathologic, molecular, and slide-level characteristics of the patient cohort are presented in Table 1. The median age was 62 years, and 57 of 100 cases (57.0%) were male. The resistant and sensitive groups were broadly similar in age, sex, smoking history, specimen type, clinical stage, predominant histologic pattern, and number of derived H&E patches (all p > 0.05).$p > 0.05$ Of the 100 EGFR-mutant patients, 58 (58.0%) had exon 19 deletions and 42 (42.0%) had the L858R point mutation, with a comparable profile between the two groups. The resistant group demonstrated a higher proportion of TP53 co-mutation than the sensitive group (62.5% vs. 38.3%, p = 0.014) and a higher Ki-67 labeling index (45.2% ± 18.6% vs. 28.7% ± 15.3%, p = 0.001).$p = 0.014 \pm \pm p < 0.001$

Table 1 Baseline clinical, pathologic, molecular, and slide-level characteristics of the patient cohort

Classification performance

Table 2 presents the predictive performance of the proposed model and baseline MIL aggregation methods. The gated-attention model achieved the highest AUC of 0.891 (95% CI: 0.838–0.936), balanced accuracy of 0.821, and F1 score of 0.808.

Table 2 Classification-performance comparison across MIL aggregation methods (5-fold cross-validation)

Model	AUC	Bal. Acc.	F1	Sens.	Spec.	PPV	NPV	ΔAUC	p
Mean pooling	0.837	0.762	0.748	0.725	0.800	0.725	0.800	−0.054	<0.01
Max pooling	0.821	0.745	0.731	0.700	0.790	0.709	0.783	−0.070	<0.01
ABMIL	0.869	0.796	0.783	0.775	0.817	0.756	0.831	−0.022	0.048
CLAM	0.874	0.801	0.789	0.780	0.822	0.762	0.836	−0.017	0.062
TransMIL	0.863	0.789	0.776	0.765	0.813	0.749	0.825	−0.028	0.035
Gated attention (ours)	**0.891**	**0.821**	**0.808**	**0.800**	**0.842**	**0.790**	**0.849**	–	–

Sens., sensitivity; Spec., specificity; PPV, positive predictive value; NPV, negative predictive value. Values represent 5-fold cross-validation averages; the confusion matrix in Fig. 4 reports pooled case numbers and corresponding row-normalized proportions.

ΔAUC is calculated relative to the proposed gated-attention model. Pairwise p values were obtained from DeLong test.

The comparative pattern across baseline models is presented graphically in Fig. 2, where the proposed framework demonstrates the highest AUC and F1 among the assessed aggregation methods.

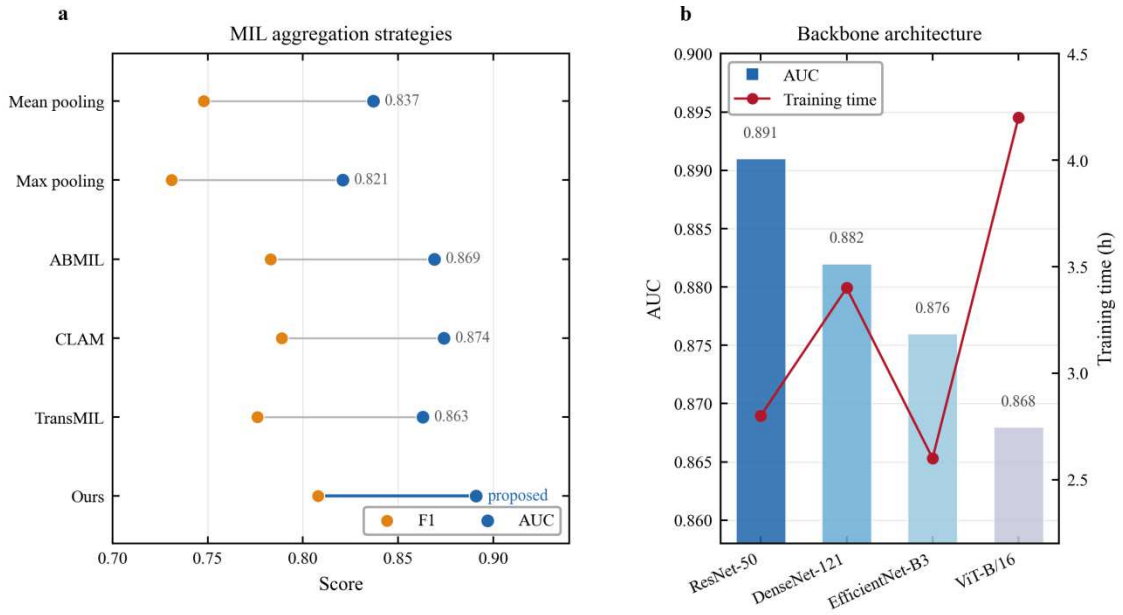


Fig. 2 Visual comparison of predictive performance. **(a)** AUC and F1 scores across MIL aggregation methods, with the proposed gated-attention model achieving the highest values. **(b)** Backbone-architecture comparison showing AUC together with training time.

DeLong test verified that the proposed framework significantly outperformed mean pooling and max pooling ($p < 0.01$) and ABMIL ($p = 0.048$). The comparison with CLAM was not statistically significant ($p = 0.062$). Class-specific performance is reported in Table 3,including ROC curves, precision-recall profiles, sample-level predicted probability distribution and model calibration results (Fig. 3).

Table 3 Class-specific diagnostic performance of the proposed framework

Class	n	AUC (95% CI)	Sensitivity	Specificity	PPV	NPV	F1
Resistant	40	0.891 (0.838–0.936)	0.800	0.850	0.780	0.864	0.790
Sensitive	60	0.891 (0.838–0.936)	0.850	0.800	0.864	0.780	0.857

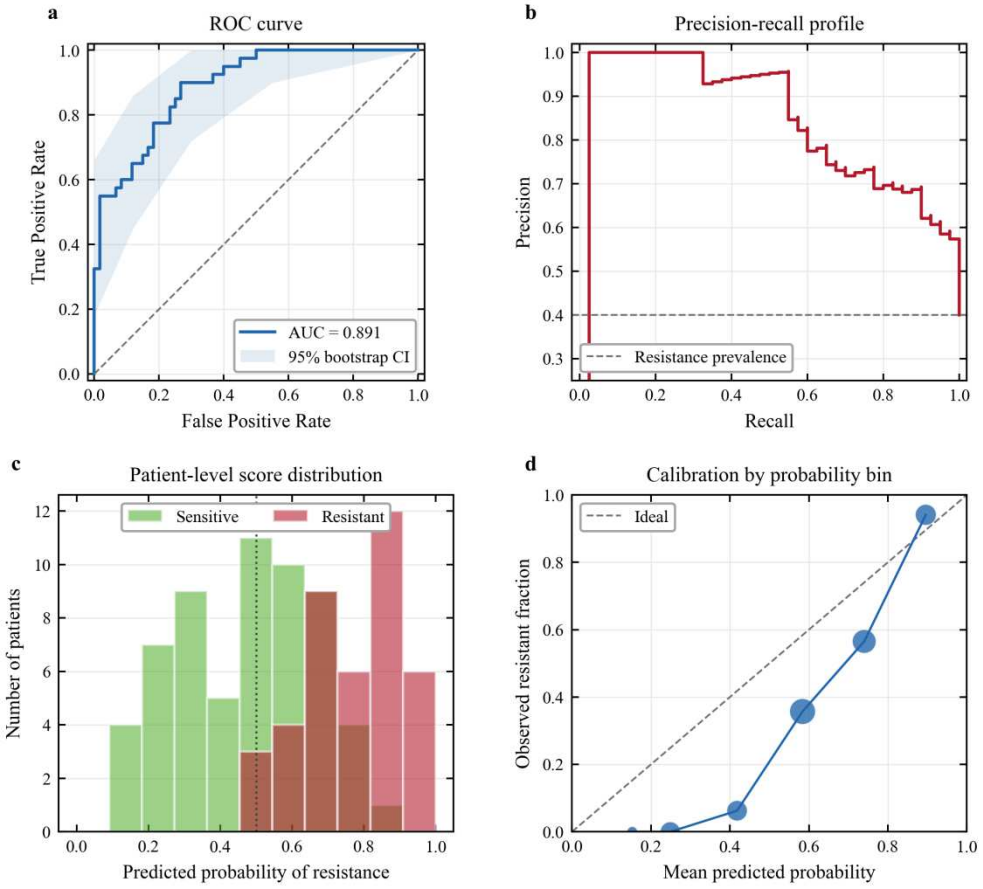


Fig. 3 Patient-level prediction performance of the proposed gated-attention model. **(a)** Receiver-operating-characteristic curve for resistant-versus-sensitive classification, with the shaded region showing the 95% bootstrap confidence interval. **(b)** Precision-recall profile, with the dashed line indicating the prevalence of resistance. **(c)** Distribution of

predicted resistance probabilities by true class. **(d)** Calibration plot comparing mean estimated probability with the observed resistance fraction across probability bins.

Confusion-matrix analysis

The confusion matrix (Fig. 4) indicated that 32 of 40 resistant cases (80.0%) and 51 of 60 sensitive cases (85.0%) were correctly classified. Eight resistant cases were incorrectly classified as sensitive, and nine sensitive cases were incorrectly classified as resistant.

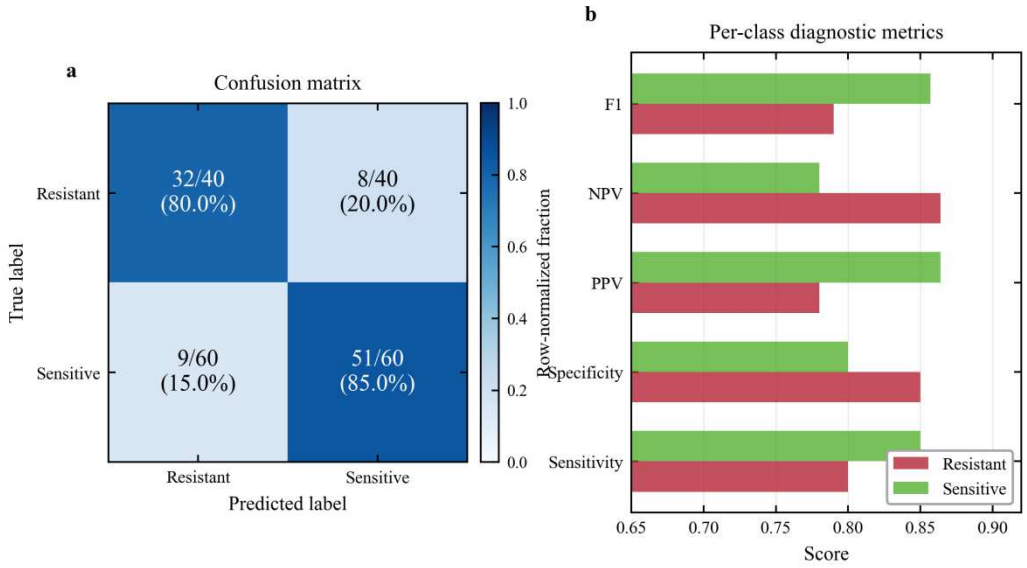


Fig. 4 Diagnostic classification summary. **(a)** Confusion matrix showing case numbers and row-normalized proportions for resistant and sensitive patients. **(b)** Per-class diagnostic metrics, including sensitivity, specificity, positive predictive value, negative predictive value, and F1 score.

Backbone-architecture comparison

Table 4 summarizes the effect of backbone architecture on predictive performance. ResNet-50 achieved the most favorable balance between performance and computational cost. DenseNet-121 yielded a comparable AUC (0.882) but required a longer training

time due to memory-intensive dense connections. EfficientNet-B3 performed slightly worse (AUC 0.876) despite its parameter efficiency. ViT-B/16 achieved an AUC of 0.868, which may be attributable to the relatively small sample size limiting the ability of Transformer-based architectures to learn discriminative patch-level representations without large-scale pathology-specific pretraining.

Table 4 Effect of backbone architecture on predictive performance

Backbone	Parameters (M)	AUC	Balanced Acc.	F1	Training time (h)
ResNet-50	23.5	**0.891**	**0.821**	**0.808**	2.8
DenseNet-121	7.0	0.882	0.809	0.796	3.4
EfficientNet-B3	10.7	0.876	0.798	0.784	2.6
ViT-B/16	85.8	0.868	0.786	0.771	4.2

Ablation analysis

To evaluate the importance of individual components, we performed ablation experiments (Table 5). Replacing gated-attention pooling with mean pooling caused the largest performance decrease (AUC $-$ 0.054), confirming that detecting the most discriminative patches is essential because of the heterogeneous nature of histopathologic specimens. Removing class-weight balancing resulted in a 2.2-point AUC decrease, mainly affecting the underrepresented resistant class. Color normalization and data augmentation each provided incremental improvements. Using all accessible patches instead of random sampling during training slightly reduced performance (AUC $-$0.008), indicating that random patch sampling serves as a regularizer.

Table 5 Ablation analysis results

Configuration	AUC	Balanced Acc.	Δ AUC
Full model (proposed)	0.891	0.821	–
− Gated attention (use mean pooling)	0.837	0.762	−0.054
− Class-weight balancing	0.869	0.793	−0.022
− Color normalization	0.876	0.801	−0.015
− Data augmentation	0.879	0.805	−0.012
− Random patch sampling (use all)	0.883	0.811	−0.008

Fig 5 provides a concise view of the performance decrease associated with each omitted component.

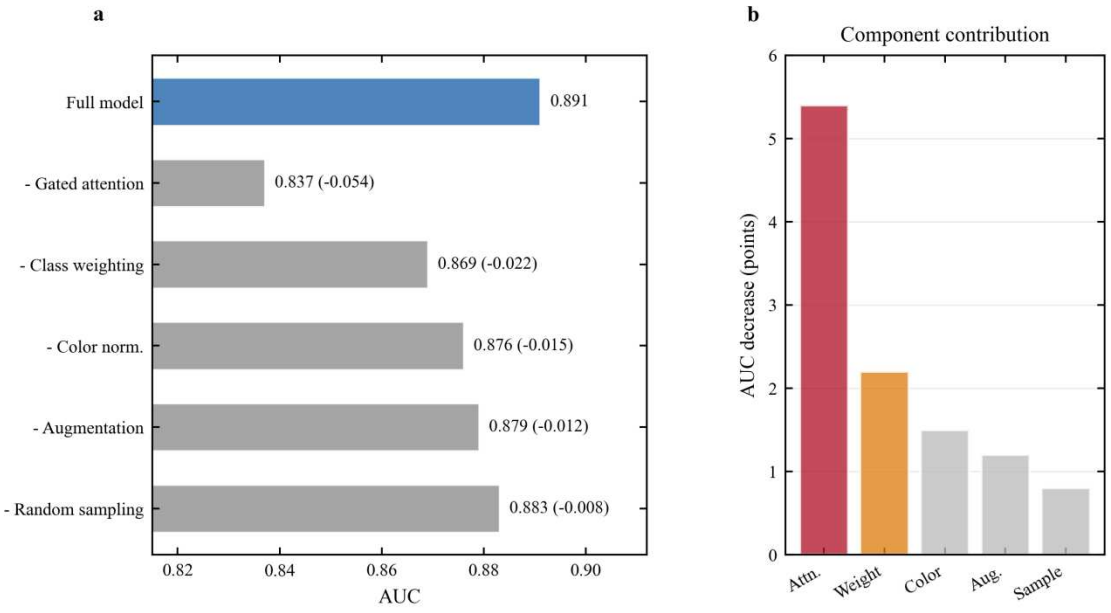


Fig. 5. Ablation analysis of the proposed model. **(a)** AUC after removing or modifying individual components; values in parentheses denote the AUC decrease relative to the full

model. **(b)** Component-wise AUC reduction, highlighting the largest contribution from gated-attention pooling.

Visualization and interpretation

Model-derived attention heatmaps (Fig. 6) identified distinct high-attention H&E areas between the two classes. In resistant cases, high-attention areas visually corresponded with areas with increased tumor-cell density and more irregular nuclear morphology, whereas sensitive cases demonstrated distinct attention patterns within the representative H&E patches. These visualizations were used as interpretability aids rather than comprehensive manual pathologic annotations.

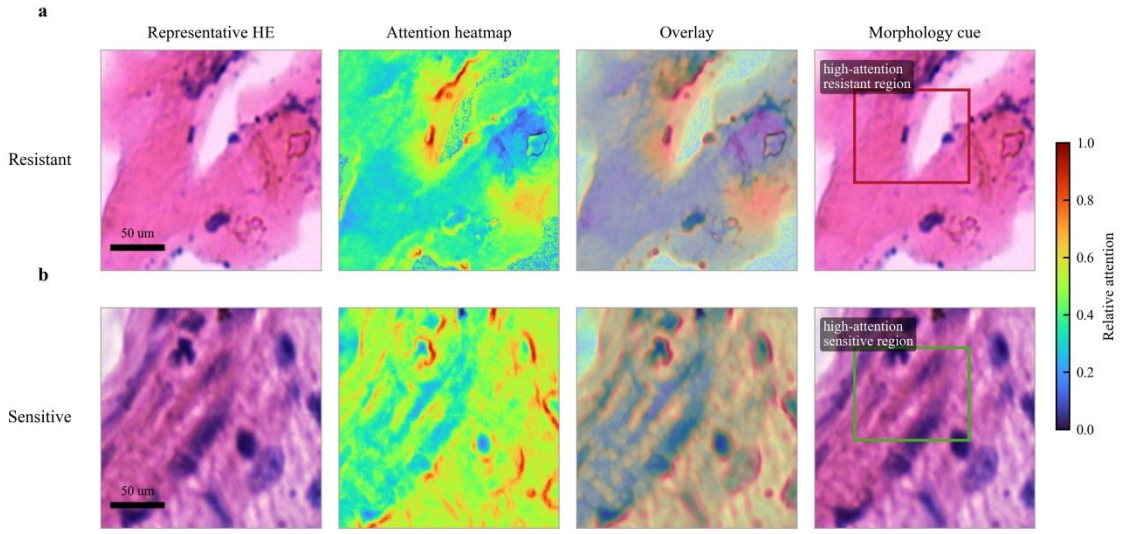


Fig. 6 Representative model-interpretability examples for EGFR-TKI resistance prediction. Rows indicate resistant and sensitive cases, respectively. Columns show the H&E patch, attention heatmap, heatmap overlay, and conservatively annotated high-attention areas. These examples demonstrate that the model assigns distinct attention patterns across resistant and sensitive cases without suggesting exhaustive manual histopathologic annotation of the selected regions.

The t-SNE projection of trained slide-level representations (Fig. 7) demonstrated partial separation between resistant and sensitive clusters, with overlap localized in cases with borderline treatment responses. This finding is consistent with the clinical reality of a continuum between sensitivity and resistance.

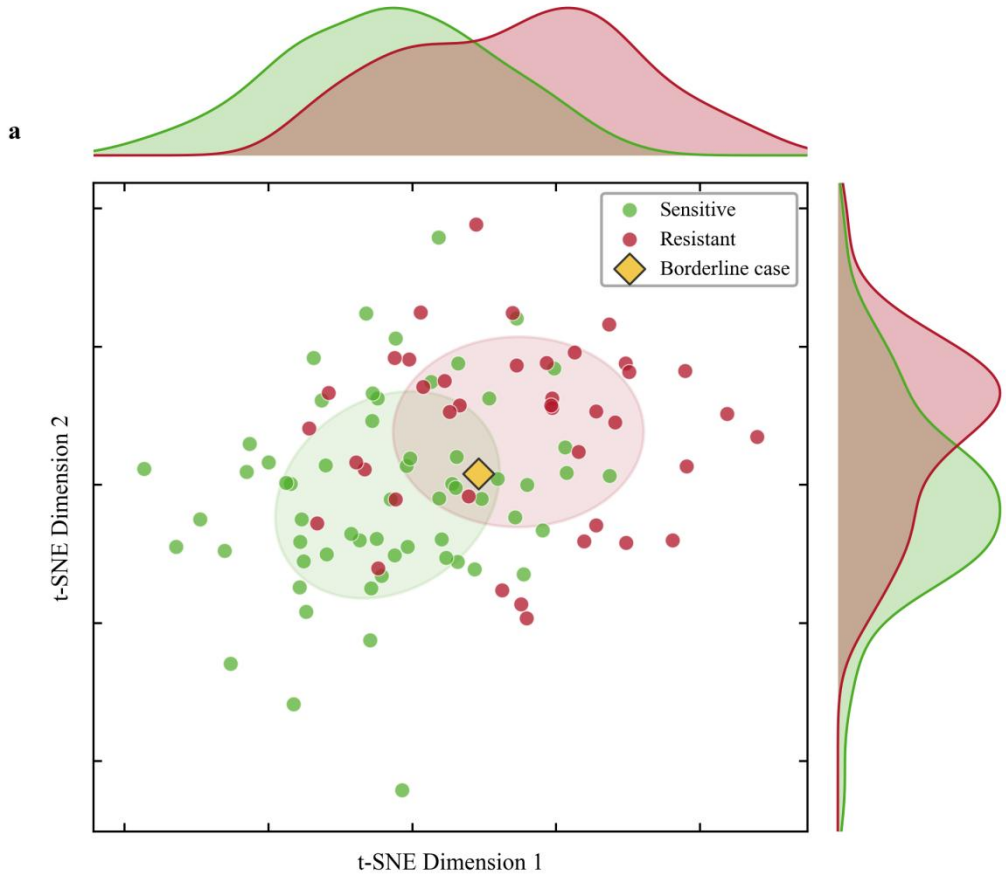


Fig.7 t-SNE projection of slide-level feature representations (z) from all patients across test folds, with marginal density distributions along each embedding dimension. Each circular point denotes one patient, colored according to the true class label; the highlighted diamond indicates a representative borderline case in the overlap region. The two groups demonstrate partial separation, indicating that the learned representation

captures resistance-associated morphology while retaining overlap compatible with clinically borderline responses.

Nuclear morphometric analysis

To characterize morphologic features associated with model-based estimations, we compared nuclear morphometric features from high-attention H&E patches between resistant and sensitive cases with HoVer-Net segmentation (Table 6; Fig. 8). Resistant cases exhibited greater nuclear sizes (78.3 ± 22.1 μm² vs. 62.7 ± 18.4 μm², p = 0.003)$\pm\mu^2 \pm \mu^2 p = 0.003$, higher nuclear-to-cytoplasmic (N/C) ratios (0.71 ± 0.14 vs. 0.58 ± 0.12, p = 0.001)$\pm \pm p < 0.001$, more irregular nuclear outlines reflected by eccentricity (0.82 ± 0.09 vs. 0.74 ± 0.11, p = 0.008)$\pm \pm p = 0.008$, and higher chromatin-texture entropy (5.84 ± 0.71 vs. 5.21 ± 0.65, p = 0.001)$\pm \pm p < 0.001$. Chromatin-texture entropy demonstrated a moderate positive association with the model-predicted resistance probability (r = 0.58, p < 0.001)$r = 0.58 p < 0.001$, indicating that nuclear texture features contribute to the trained decision boundary.

Table 6 Nuclear morphometric features in high-attention H&E areas

Feature	Resistant	Sensitive	Difference	pP-value
Nuclear area (μm²)μ^2	78.3 ± 22.1±	62.7 ± 18.4±	+15.6	0.003
Nuclear-to-cytoplasmic ratio	0.71 ± 0.14±	0.58 ± 0.12±	+0.13	<0.001
Nuclear eccentricity	0.82 ± 0.09±	0.74 ± 0.11±	+0.08	0.008
Chromatin texture entropy	5.84 ± 0.71±	5.21 ± 0.65±	+0.63	<0.001

Values are mean ± SD± from high-attention H&E patches segmented with HoVer-Net. Comparisons were made between resistant and sensitive cases using the Mann-Whitney U U test.

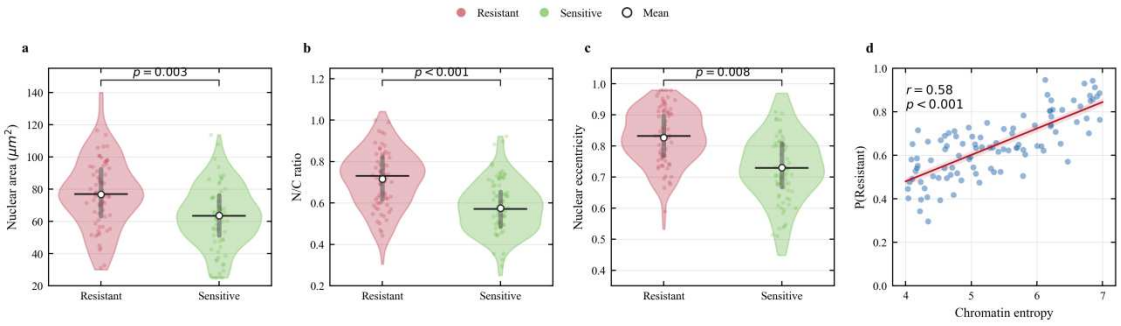


Fig. 8. Nuclear morphometric analysis of high-attention H&E patches. **(a)** Nuclear area in resistant and sensitive cases (p = 0.003).$p = 0.003$ **(b)** Nuclear-to-cytoplasmic ratio (p = 0.001).$p < 0.001$ **(c)** Nuclear eccentricity (p = 0.008).$p = 0.008$ **(d)** Chromatin texture entropy versus model-predicted resistance probability, showing a positive association (r = 0.58, p < 0.001).$r = 0.58 p < 0.001$

Subgroup analysis by EGFR mutation type

We evaluated whether predictive performance differed between cases with exon 19 deletions and L858R mutations (Table 7). The framework yielded an AUC of 0.903 (95% CI: 0.847–0.949) among exon 19 deletion carriers and 0.871 (95% CI: 0.798–0.932) among L858R carriers. The modestly lower performance in the L858R subgroup may indicate greater phenotypic variability in this mutation subgroup, as L858R-mutant tumors exhibit more variable histologic patterns[4].

Table 7 Subgroup analysis by EGFR mutation type

Subgroup	n (R/S)	AUC (95% CI)	Sensitivity	Specificity
Exon 19 deletion	58 (22/36)	0.903 (0.847–0.949)	0.818	0.889
L858R	42 (18/24)	0.871 (0.798–0.932)	0.778	0.833
All	100 (40/60)	0.891 (0.838–0.936)	0.800	0.850

Discussion

Key findings

The present study proposes a weakly supervised multiple-instance learning (MIL) pipeline capable of predicting early treatment responses directly from routinely available H&E images without requiring pixel-level annotations [13, 14]. This strategy was designed for patients with EGFR-mutant lung adenocarcinoma receiving EGFR-TKI therapy [1, 2]. In a retrospective cohort of 100 patients, the framework achieved an AUC of 0.891, indicating that routine H&E WSIs contain consistent morphologic signals that can be effectively detected by deep-learning models [10, 33, 34, 9].

Notably, these signals may not correspond to readily discernible histopathologic features [10, 33], but they are consistent with a latent morphologic risk phenotype, which is encoded by spatial tumor architecture. Disease progression or poor response within 12 weeks represents a clinical endpoint that is more closely aligned with the concept of primary resistance or early treatment failure than with classical acquired resistance [3]. Consequently, the framework primarily reflects baseline intrinsic differences in drug sensitivity, rather than later clonal evolution, and may support pretreatment or early-treatment risk stratification.

Methodologically, gated-attention MIL outperformed mean pooling, max pooling, and Transformer-based models (TransMIL) [13, 15]. This suggests that in medium-sized WSI cohorts predictive performance depends more on the inductive bias introduced by explicit informative-instance selection than on global self-attention modeling [15]. The ablation experiments also demonstrated that using all patches unexpectedly reduced performance, likely because of increased noise, reduced signal specificity, and attention dilution as bag size increased. Such findings reinforce the MIL principle that predictions are driven by a constrained number of discriminative tissue instances [13].

Mechanistic interpretation

From biological and computational perspectives, interpretability analyses demonstrated substantial concordance between Grad-CAM heatmaps and HoVer-Net-based cellular segmentation, where the framework focused primarily on regions with nuclear atypia, increased cellular density, and architectural disorganization [25, 27]. The quantitative morphometric analysis further indicated that resistant/poor-response tumors exhibited a greater nuclear size, higher nuclear-to-cytoplasmic ratio, and more irregular nuclear contours [27], indicating greater spatial variability and tissue-level remodeling.

At the genomic level, TP53 co-mutation was also markedly enriched in the resistant/poor-response group (62.5% vs. 38.3%, $p = 0.014$), which is consistent with the previous evidence linking TP53 co-mutation, genomic instability, and poor outcomes in lung adenocarcinoma [32, 33, 9]. Mechanistically, inactivation of TP53 may impair DNA repair and promote chromosomal instability, which may result in remodeling of nuclear architecture and enhanced spatial variability. These downstream morphologic alterations

are captured by the MIL framework as a latent morphologic risk phenotype, which is associated with lower EGFR-TKI sensitivity. This finding suggests that deep-learning models do not directly detect molecular alterations, but rather infer genomic instability from histopathologic phenotypes [10, 33]. It may be considered as phenotype-mediated learning in which tissue-level spatial organization serves as an intermediate representation of latent molecular states. The results of subgroup analysis also showed superior performance in EGFR exon 19 deletion (19del) than in L858R mutations (AUC = 0.903 and 0.871) [34], which may indicate differences in tumor-microenvironment complexity and spatial heterogeneity among EGFR subtypes .

Clinical and methodological implications

Although the predictive performance and mechanistic interpretability were encouraging, the present study had some limitations [9, 10]. There was no independent external validation cohort, which may limit generalizability; therefore, it is essential to conduct multicenter validation and evaluation using publicly accessible datasets, such as TCGA-LUAD [30]. Nevertheless, recent pathology foundation models, including Virchow, UNI, and CONCH, which are pretrained on large-scale histopathology datasets through self-supervised learning, provide strong cross-domain generalization [21–23].

This supports a more scalable design framework, where a frozen foundation-model backbone extracts features, and is combined with a lightweight gated-attention MIL classification head to improve robustness across institutions and datasets [21, 22]. In addition, the lack of stratification by EGFR-TKI generation may represent an important

confounder because different TKI generations have distinct pharmacologic effects and resistance landscapes that should be addressed in future multivariable models [2, 3].

Because longitudinal histologic and paired molecular data were unavailable, we could not directly establish causal links between morphologic patterns and specific resistance mechanisms (such as EGFR T790M or MET amplification) [3, 5]. Therefore, the current framework should be interpreted as a therapy-related morphologic risk model rather than a mechanistic model of molecular resistance pathways.

Overall, this study demonstrates the feasibility of lightweight MIL-based computational pathology models and provides preliminary evidence that multiscale modeling can link genomic instability, tumor spatial heterogeneity, and treatment response [33].

Conclusion

In summary, we developed a weakly supervised MIL-based computational-pathology model capable of predicting early treatment response in EGFR-mutant lung adenocarcinoma with standard H&E-stained WSIs alone. The model identifies latent morphologic patterns associated with tumor spatial variability and shows robust predictive performance without pixel-level annotations. Biologically, these phenotypes may indicate underlying genomic instability, particularly TP53-associated chromosomal dysregulation, which is manifested as nuclear architectural remodeling and spatial heterogeneity. Clinically, this framework may enable early risk stratification before or at initiation of EGFR-TKI therapy and underscores the potential for integrating weakly

supervised learning with foundation models to build scalable and explainable digital-pathology systems.

Detailed model hyperparameters

Table 8 lists the hyperparameters used in the final model configuration.

Table 8 Model hyperparameters

Abbreviations

Declarations

Funding

This study was supported by the Natural Science Foundation of Hebei Province (Grant No. H2023106015).

Availability of data and materials

The clinical and pathological summary data supporting the conclusions of this article are included within the text. Following the protocol approved by the Institutional Review Boards, the raw whole-slide images (WSIs) and the deep learning implementation codes are not publicly hosted in a repository due to institutional data protection policies. However, they can be made available from the corresponding author upon reasonable academic request for replication or peer-review purposes.

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and relevant national ethical guidelines, and was approved by the Institutional Review Board of

Shijiazhuang People's Hospital and Hengshui People's Hospital (Approval Numbers: 2023-KY-24 and 2026-KY-105). The requirement for informed consent was waived by the Institutional Review Board due to the retrospective nature of this study.**Consent for publication**

Not applicable.

Authors' contributions

W.W. and P.Z. contributed to study conception, digital pathology data acquisition, image preprocessing, clinical cohort curation, deep learning model development, statistical analysis, and visualization. W.W. drafted the manuscript. Y.Z. supervised the overall study design, ethical compliance, and manuscript revision. All authors reviewed and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge with gratitude all clinical, pathological, and technical staff members from Shijiazhuang People's Hospital and Hengshui People's Hospital who participated in or contributed to this study.

Competing interests

The authors declare that they have no conflict of interest.

References

[1] Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med. 2009;361:947-57. doi:10.1056/NEJMoa0810699.

[2] Soria JC, Ohe Y, Vansteenkiste J, Reungwetwattana T, Chewaskulyong B, Lee KH, et al. Osimertinib in untreated EGFR-mutated advanced non-small-cell lung cancer. N Engl J Med. 2018;378:113-25. doi:10.1056/NEJMoa1713137.

[3] Oxnard GR, Hu Y, Mileham KF, Husain H, Costa DB, Tracy P, et al. Assessment of resistance mechanisms in EGFR T790M-positive lung cancer and acquired resistance to osimertinib. JAMA Oncol. 2018;4:1527-34. doi:10.1001/jamaoncol.2018.2969.

[4] Zhang Y, Sheng J, Kang S, Fang W, Yan Y, Hu Z, et al. Patients with exon 19. deletion were associated with longer progression-free survival compared to those with L858R mutation after first-line EGFR-TKIs. PLoS One. 2014;9:e107161. doi:10.1371/journal.pone.0107161.

[5] Zhao J, Xu W, Zhou F, Zhang X, Zhou M, Miao D, et al. Navigating the landscape of EGFR TKI resistance in EGFR-mutant NSCLC — mechanisms and evolving treatment approaches. Nat Rev Clin Oncol. 2026;23:63–83. doi:10.1038/s41571-025-01085-z.

[6] Sacher AG, Paweletz C, Dahlberg SE, Alden RS, O'Connell A, Feeney N, et al. Prospective validation of rapid plasma genotyping for the detection of EGFR and KRAS mutations in advanced lung cancer. JAMA Oncol. 2016;2:101414. doi:1001/jamaoncol.2016.0173.

[7] Pennell NA, Arcila ME, Gandara DR, West H. Biomarker testing for patients with advanced non-small cell lung cancer: real-world issues and tough choices. ASCO Educ Book. 2019;39:531–42. doi:10.1200/EDBK_237863.

[8] Lindeman NI, Cagle PT, Aisner DL, Arcila ME, Beasley MB, Bernicker EH, et al. Updated molecular testing guideline for the selection of lung cancer patients for

treatment with targeted tyrosine kinase inhibitors. J Mol Diagn. 2018;20:129–59. doi:10.1016/j.jmoldx.2017.11.004.

[9] Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. 2019;20:e253–61. doi:10.1016/S1470-2045(19)30154-8.

[10] Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020;11:3877. doi:10.1038/s41467-020-17678-4.

[11] Juan RA, Parmar C, Carrasco-Zevallos OM, Csiszer C, Yip SSF, Raciti P, et al. Development and deployment of a histopathology-based deep learning algorithm for patient prescreening in a clinical trial. Nat Commun. 2024;15:4690. doi: 10.1038/s41467-024-49153-9.

[12] Pao JJ, Biggs M, Duncan D, Lin DI, Davis R, Huang RSP, et al. Predicting EGFR mutational status from pathology images using a real-world dataset. Sci Rep. 2023;13:4404. doi: 10.1038/s41598-023-31284-6.

[13] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML). 2018;80:2127-36. doi:10.48550/ARXIV.1802.04712.

[14] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng. 2021;5:555-70. doi:10.1038/s41551-020-00682-w.

[15] Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X, et al. TransMIL: transformer-based correlated multiple instance learning for whole slide image classification. In:

Advances in Neural Information Processing Systems (NeurIPS). 2021;34:2136-47. doi:10.48550/arXiv.2106.00908

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770-8. doi:10.1109/CVPR.2016.90.

[17] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:4700-8. doi:10.1109/CVPR.2017.243.

[18] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). 2019;97:6105-14. doi:10.48550/arXiv.1905.11946.

[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). 2017;30:5998-6008. doi:10.48550/arXiv.1706.03762.

[20] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). 2021. doi:10.48550/arXiv.2010.11929.

[21] Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30:1353-63. doi:10.1038/s41591-024-02857-3.

[22] Xu H, Usuyama N, Zhang S, Rao J, Li J, Naumann T, et al. A whole-slide foundation model for digital pathology from real-world data. Nature. 2024;630:181–8. doi:10.1038/s41586-024-07441-w.

[23] Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, et al. A vision-language foundation model for computational pathology. Nature. 2024;634:466–73. doi:10.1038/s41586-024-07618-3.

[24] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44:837. doi:10.2307/2531595.

[25] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017:618-26. doi:10.1109/ICCV.2017.74.

[26] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.

[27] Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. HoVer-Net: simultaneous segmentation and classification of nuclei in histology images. Med Image Anal. 2019;58:101563. doi:10.1016/j.media.2019.101563.

[28] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009:248-55. doi:10.1109/CVPR.2009.5206848.

[29] Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). 2009:1107-10. doi:10.1109/ISBI.2009.5193250.

[30] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020 : GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-49. doi:10.3322/caac.21660.

[31] Travis WD, Brambilla E, Nicholson AG, Yatabe Y, Austin JHM, Beasley MB, et al. The 2015 WHO classification of lung tumors : impact of genetic, clinical and radiologic advances since the 2004 classification. J Thorac Oncol. 2015;10(9):1243-60. doi:10.1097/JTO.0000000000000630.

[32] Dong ZY, Zhong WZ, Zhang XC, Su J, Xie Z, Liu SY, et al. Potential predictive value of TP53 and KRAS mutation status for response to PD-1 blockade immunotherapy in lung adenocarcinoma. Clin Cancer Res. 2017;23:3012-24. doi:10.1158/1078-0432.CCR-16-2554.

[33] Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat Cancer. 2020;1:800-10. doi:10.1038/s43018-020-0085-5.

[34] Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24:1559-67. doi:10.1038/s41591-018-0177-5.

Table 1 Baseline clinical, pathologic, molecular, and slide-level characteristics of the patient cohort

Characteristic	Resistant (n = 40)$n = 40$	Sensitive (n = 60)$n = 60$	p P-value
Age, median (range), years	(38–77)	(41–79)	
Male, n (%)	(60.0)	(55.0)	
Ever-smoking history, n (%)	(45.0)	(36.7)	
Specimen type			
Surgical resection	(77.5)	(78.3)	
Biopsy	(22.5)	(21.7)	
Tumor stage			
I	(15.0)	(20.0)	
II	(20.0)	(23.3)	
III	(35.0)	(33.3)	
IV	(30.0)	(23.3)	
Predominant histological pattern			
Acinar	(42.5)	(45.0)	
Papillary	(20.0)	(23.3)	
Lepidic	(7.5)	(8.3)	
Solid	(22.5)	(16.7)	

Characteristic	Resistant (n = 40)$n = 40$	Sensitive (n = 60)$n = 60$	p P-value
Micropapillary/other	(7.5)	(6.7)	
EGFR testing method			
PCR-based assay	(65.0)	(65.0)	
NGS	(35.0)	(35.0)	
EGFR mutation type			
Exon 19 deletion	(55.0)	(60.0)	
L858R	(45.0)	(40.0)	
First-line EGFR-TKI			
Gefitinib	(27.5)	(31.7)	
Erlotinib	(20.0)	(20.0)	
Icotinib	(22.5)	(23.3)	
Osimertinib	(30.0)	(25.0)	
TP53 co-mutation, n (%)n	(62.5)	(38.3)	0.014
Ki-67 labeling index, mean ± SD (%)±	±45.2 ± 18.6	±28.7 ± 15.3	<0.001
H&E patches per patient, mean ± SD	± 132	± 125	

Abbreviations: H&E, hematoxylin and eosin; NGS, next-generation sequencing; PCR, polymerase chain reaction; SD, standard deviation; TKI, tyrosine kinase inhibitor. Unless otherwise indicated, values are shown as n (%). Comparisons used the Mann-Whitney U U test for continuous variables and the chi-square or Fisher's exact test for categorical variables.

Table 8 Model hyperparameters

Hyperparameter	Value
Input patch size	512×512 pixels
Magnification level	20× (0.5 μμm/pixel)
Backbone architecture	ResNet-50 (ImageNet-pretrained)
Feature dimension (backbone output)	2048
Attention hidden dimension	256
MLP hidden dimension	256
Classifier output	2 (binary)
Dropout rate	0.5
Optimizer	AdamW
Initial learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Learning rate schedule	Cosine annealing with warm restarts
Batch size	16
Maximum epochs	80
Early stopping patience	15 epochs
Patches sampled per patient (training)	100