



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2026, 2026:92

<https://doi.org/10.28919/cmbn/10144>

ISSN: 2052-2541

DEEP LEARNING APPROACHES FOR NON-INVASIVE PREDICTION OF EGFR MUTATION STATUS IN LUNG ADENOCARCINOMA FROM HISTOPATHOLOGICAL AND RADIOLOGICAL IMAGES: A REVIEW

WAHYUDI SETIAWAN*, ADE VERDAUS SAPUTRA

Department of Information Systems, Universitas Trunodjoyo Madura, Bangkalan 69162, Indonesia

Copyright © 2026 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: Lung adenocarcinoma patients with an EGFR gene mutation respond well to targeted drugs called EGFR-TKIs. Testing for this mutation needs a tissue biopsy and lab work. This process is invasive, costly, and slow, and many patients do not have enough tissue for testing. This review asks whether deep learning can predict EGFR mutation status from images doctors already collect. These images include histopathology slides and radiology scans such as CT, PET/CT, or MRI. We searched Semantic Scholar, PubMed, Scopus, and arXiv through the Consensus platform in July 2026. The search covered studies published from 2018 to 2026. This search found 30 records, and 28 remained after removing duplicates. Nineteen studies met our criteria, reported across 20 publications. Seven studies used histopathology images, eleven used radiology images, and one combined both. Histopathology-based models reached accuracy scores (AUC) between 0.682 and 0.933. Radiology-based models reached AUC between 0.821 and 0.940. The best results came from models that combined feature types, such as radiomics with deep learning. Other strong results came from models built on large pretrained pathology foundation models, such as UNI, CONCH, and TITAN. One early study fused radiology and histopathology directly, pointing to a new multimodal direction. Most

*Corresponding author

E-mail address: wsetiawan@trunojoyo.ac.id

Received July 18, 2026

studies used heatmaps to explain their predictions, but few checked whether these explanations were reliable. Deep learning can predict EGFR mutation status from routine images. Current studies use small patient groups, and few test their models on new hospitals. Their explanations are not well validated. Larger, better-tested studies are still needed before this approach reaches clinical use.

Keywords: EGFR mutation; lung adenocarcinoma; deep learning; whole-slide image; radiomics.

2020 AMS Subject Classification: 92C50, 68T07.

1. INTRODUCTION

Lung cancer is the most common cancer and the leading cause of cancer death worldwide. It caused an estimated 2.48 million new cases and 1.8 million deaths in 2022 [1]. This disease continues to carry one of the highest death rates among all cancers and remains a major global health burden, prompting the World Health Organisation to prioritize early detection and equitable treatment access as part of its cancer-control agenda. Advances in biomedical imaging and computational analysis are increasingly viewed as tools to help ease this burden through earlier, non-invasive diagnosis. Adenocarcinoma is the main subtype of non-small cell lung cancer (NSCLC). Mutations in the epidermal growth factor receptor (EGFR) gene are among its most actionable alterations, mainly exon 19 deletions and the exon 21 L858R substitution.

Patients with sensitizing EGFR mutations respond well to EGFR tyrosine kinase inhibitors (EGFR-TKIs) [2]. EGFR mutation frequency differs by ethnicity: about 40-50% in Asian adenocarcinoma patients versus 10-20% in Western patients [3]. Treatment outcomes after TKI therapy also differ across ethnic groups [4]. This makes accessible, population-representative testing important.

Standard EGFR testing uses polymerase chain reaction (PCR) or next-generation sequencing (NGS) on tissue from biopsy or surgery. These methods are invasive, costly, and slow. Small tissue samples may miss tumor heterogeneity. Many patients with advanced disease lack enough tissue for testing, and results can take one to several weeks.

Deep learning offers an alternative route to predict EGFR status from images already collected

during routine care. Two imaging types dominate this research. The first is digital histopathology, where whole-slide images (WSIs) of hematoxylin and eosin (H&E)-stained tissue are analyzed using convolutional neural networks (CNNs), attention-based multiple instance learning (MIL), or pathology foundation models. Coudray et al. showed that a CNN trained on TCGA whole-slide images could classify tumor subtype and predict several mutated genes, including EGFR, from tissue shape alone [5],[6].

The second imaging type is radiology. CT, positron emission tomography/CT (PET/CT), and magnetic resonance imaging (MRI) scans are analyzed using radiomics, deep convolutional features, or fusion models to predict EGFR status without tissue sampling.

This literature is split across two modalities, many model types (CNNs, MIL variants, vision transformers, radiomics-deep learning fusion, and foundation models), and cohorts of different size and origin.

1.1 Novelty Compared with Previous Reviews

Several prior reviews cover parts of this literature, but each stays within one modality or one narrow scope. Table 1 compares these reviews with the present one.

Table 1. Comparison of prior reviews on AI-based EGFR mutation prediction with the present review.

Prior Review (Year)	Modality Scope	Genes Covered	Studies (n)	Foundation Models	Multimodal Fusion
Ge et al. (2023) [7]	PET/CT only	EGFR	Narrative (no fixed n)	Not discussed	Not discussed
Ma et al. (2024) [8]	PET/CT only	EGFR	17	Not discussed	Not discussed
Nguyen et al. (2024) [9]	CT / PET radiomics only	EGFR	Searched to Feb 2022	Not covered (predates)	Not discussed
Fuster-Matanzo et al. (2025) [10]	CT radiomics only	EGFR, ALK, KRAS	124	Not discussed	Not discussed
Parra-Medina et al. (2025) [11]	Histopathology only	EGFR, ALK, ROS1, KRAS, TP53, TMB	Not stated here	Some studies included	Not discussed
This review	Histopathology radiology + multimodal	+EGFR	19	UNI, CONCH, TITAN	Yes (1 study)

Four prior reviews cover only radiology. Ge et al., Ma et al., and Nguyen et al. cover CT or PET/CT radiomics alone [7], [8], [9]. Fuster-Matanzo et al. also cover CT radiomics alone, across three genes rather than EGFR specifically [10]. None of these four reviews include histopathology-based prediction.

Parra-Medina et al. cover histopathology-based prediction, but across six molecular targets, not EGFR specifically [11]. None of these five reviews discuss pathology foundation models, such as UNI, CONCH, or TITAN, in a systematic way. None discuss fusing radiology and histopathology in one model.

This review is the first to combine histopathology- and radiology-based EGFR prediction in one synthesis. It is also the first to connect this literature to pathology foundation models and multimodal fusion. This addresses a real gap, rather than repeating prior work. It answers three questions:

RQ1. What deep learning architectures have been used to predict EGFR mutation status from histopathology and radiology images of lung adenocarcinoma, and how does their performance compare? (Addressed in Sections 3.2-3.5 and in Section 4.)

RQ2. What interpretability and validation methods support the clinical credibility of these models? (Addressed in Sections 3.2-3.4, 3.6, and in Section 4.)

RQ3. What gaps and barriers must be addressed before image-based EGFR prediction can move toward clinical use? (Addressed in Section 4, Section 5, and Section 6.)

2. MATERIALS AND METHODS

2.1 Protocol and Reporting Standard

This review followed the PRISMA 2020 statement [12]. The scope was adapted for a review suited to a focused technology synthesis. The review was not registered in advance. Section 5 explains this and other limits.

2.2 Search Strategy

Literature was searched across Semantic Scholar, PubMed, Scopus, and arXiv using the Consensus platform in July 2026. Two initial queries covered histopathology-based and radiology-based EGFR prediction. A third, supplementary query covered multimodal fusion of radiology and histopathology. Citation snowballing from key articles was also used to cross-check completeness. Search terms combined four concepts: (i) EGFR mutation, (ii) lung adenocarcinoma or NSCLC,

(iii) deep learning, CNN, MIL, radiomics, or foundation model, and (iv) whole-slide image, histopathology, CT, PET/CT, or MRI. Only English-language studies were searched. The retrieved studies span 2018 to 2026, following the introduction of large digitized WSI repositories such as The Cancer Genome Atlas (TCGA).

2.3 Eligibility Criteria

This review used the PICOS framework (Population, Intervention, Comparator, Outcomes, Study Design) to define eligibility. Table 2 summarizes the PICOS criteria.

Table 2. PICOS criteria used to define study eligibility.

Element	Criterion
Population	Patients with lung adenocarcinoma, or adenocarcinoma-predominant non-small cell lung cancer (NSCLC), of any age, sex, or ethnicity.
Intervention (Index Test)	A deep learning model, or a model combining deep learning with radiomics, applied to H&E-stained whole-slide histopathology images or to CT, 18F-FDG PET/CT, or MRI scans, to predict EGFR mutation status.
Comparator (Reference Standard)	Molecular EGFR genotyping by polymerase chain reaction (PCR) or next-generation sequencing (NGS), used as the ground-truth label for model training or validation.
Outcomes	Primary: a quantitative diagnostic performance metric for EGFR mutation status (AUC, accuracy, sensitivity, or specificity). Secondary: use of an interpretability method; internal versus external validation.
Study Design	Retrospective diagnostic-accuracy or prediction-model development studies, including conference abstracts with an extractable metric. Narrative reviews, meta-analyses without new primary data, and studies predicting a different outcome (e.g., treatment response) were excluded.

Studies were included if they met four conditions, consistent with the PICOS criteria above. First, they used or built a deep learning model, or a model combining deep learning with radiomics. Second, they used H&E whole-slide histopathology images or CT/PET/MRI scans as input. Third, they predicted EGFR mutation status in lung adenocarcinoma or adenocarcinoma-predominant NSCLC. Fourth, they reported a quantitative metric such as AUC, accuracy, or sensitivity/specificity.

Studies were excluded if EGFR prediction was only a secondary analysis without a clear metric. Studies using only classical machine learning, with no deep-learning part, were also excluded. Studies reporting only other genes, such as ALK or KRAS, without separate EGFR results were excluded, as were narrative commentaries without original data. Studies predicting a different outcome, such as treatment response in already-confirmed EGFR-mutant patients, were also excluded, since the outcome of interest here is mutation status itself. Studies reporting only a single

overall accuracy value, without AUC, sensitivity, or specificity, were excluded as non-comparable for this synthesis. Conference abstracts were included only if they reported an extractable metric. These are flagged in Table 3.

2.4 Study Selection and Data Extraction

Titles and abstracts were screened first. Full texts or full abstracts were then checked against the eligibility criteria. For each included study, the following data were extracted: author, year, imaging modality, data source, cohort size, model architecture, validation type (internal or external), best reported AUC, and whether an interpretability method was used. Figure 1 shows the resulting PRISMA 2020 flow diagram.

After initial extraction, each record's key details were checked again against the original abstract or full text. These details included author names, year, journal, sample size, and best reported AUC. This step caught and corrected several transcription errors during preparation of this draft.

Nineteen primary studies met the criteria: seven histopathology-based, eleven radiology-based, and one multimodal fusion study combining radiology and histopathology, reported across twenty publications. Three systematic reviews and meta-analyses on related oncogene prediction were found in the same searches [7],[8],[10]. These were kept as reference syntheses, not as primary data. Supporting literature on MIL architectures, foundation models, and explainable AI was found through separate searches. This literature supports the discussion in Sections 3.5 and 3.6.

Risk of bias for each included study was assessed using the QUADAS-2 tool across its four domains (patient selection, index test, reference standard, and flow and timing), rated as low, some concerns, or high risk. One reviewer performed the assessment; results are summarized by modality in Table 4.

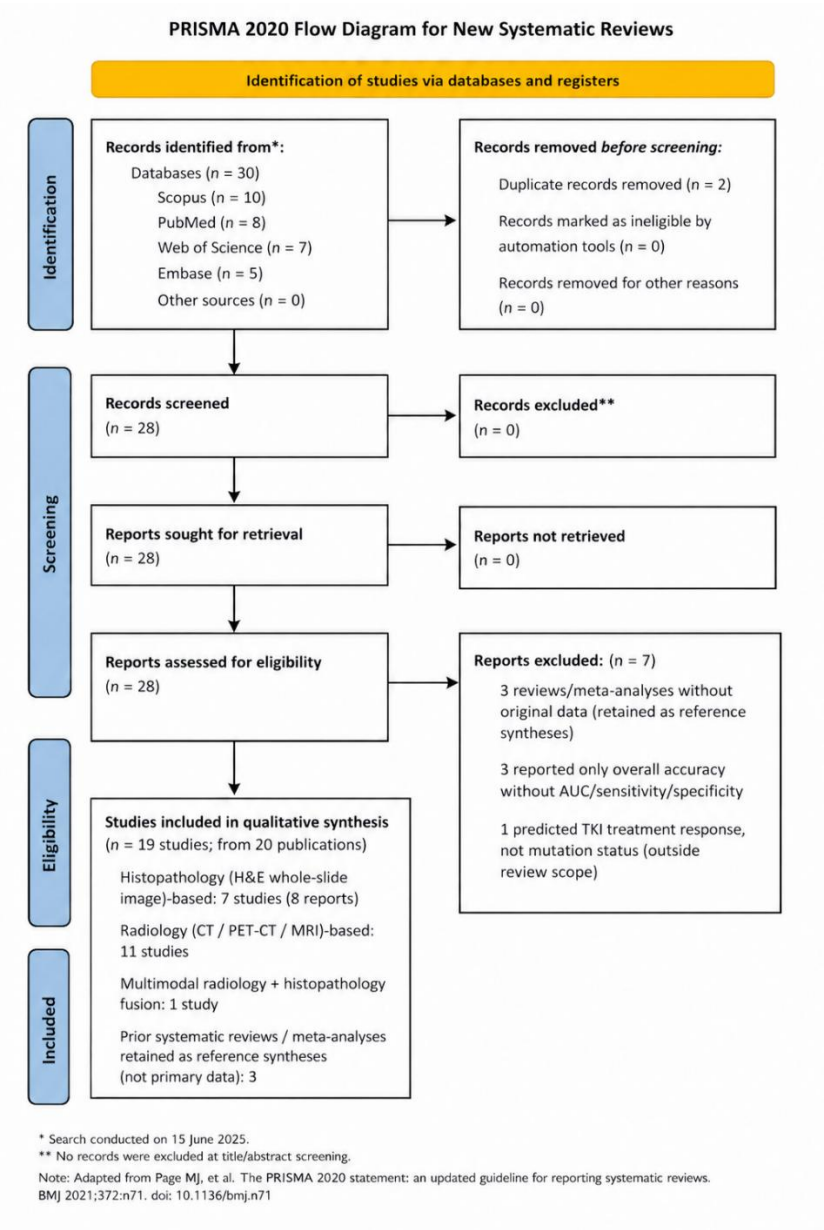


Figure 1. PRISMA 2020 flow diagram of the study identification, screening, and inclusion process.

3. RESULTS

3.1 Overview of Included Studies

Figure 2 shows the two main pipelines covered in this review. It orients readers before the study-by-study findings that follow.

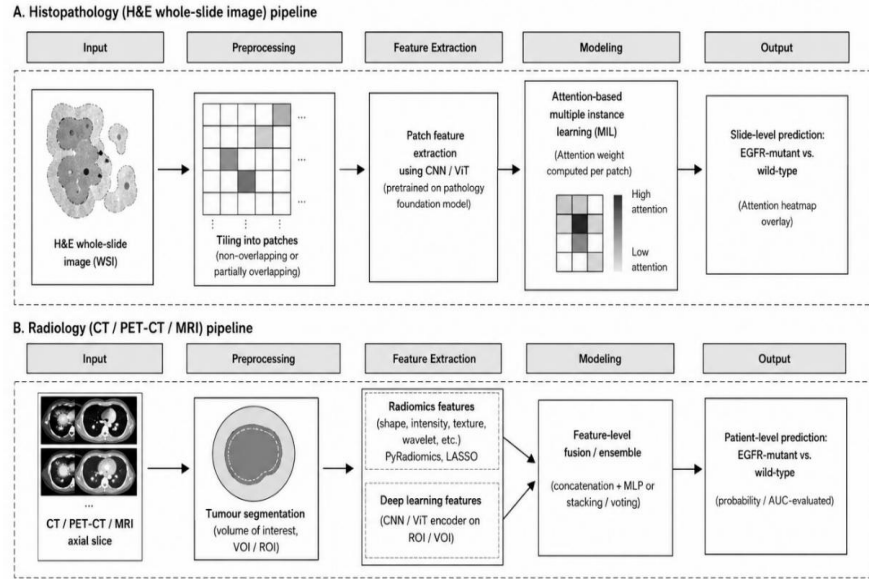


Figure 2. Schematic overview of the histopathology-based (Panel A) and radiology-based (Panel B) EGFR-prediction pipelines synthesised in this review. Illustrative schematic; not derived from or reproducing any specific patient image from the reviewed studies.

Table 3. Summary of primary studies applying deep learning to EGFR mutation prediction in lung adenocarcinoma from histopathological or radiological images.

Study (Year)	Modality	Data Source / Cohort	Sample Size	Model / Architecture	Best Reported AUC
Coudray et al. (2018) [5],[6]	H&E WSI	TCGA-LUAD	~1,600 slides	Inception-v3 (tile-level CNN)	0.826 (EGFR)
Tomita et al. (2022) [13]	H&E WSI (FFPE)	Dartmouth-Hitchcock CPTAC-3	747 WSIs / 232 patients	CNN, weakly supervised	0.799 (internal); 0.686 (external)
Park J.H. et al. (2024) [14]	H&E WSI	3 Korean hospitals + Ajou (test)	868 train / 197 test WSIs	MIL with patch-masking scheduler	0.768 (AUROC)
Chen C.L. et al. (2025) [15]†	H&E WSI	Single-center (Taiwan)	1,572 WSIs	Weakly supervised AI (annotation-free)	0.682 (same-site); 0.656 (cross-site)
Shao et al. (2026) [16]	H&E WSI	Single-center retrospective	268 patients	Attention-based MIL + DenseNet201 + logistic regression	0.885
Gwal et al. (2025) [17]	H&E WSI	Indian cohort + TCGA (external)	170 train / 30 internal / 86 external	ViT foundation model + ABMIL	0.933 (internal); 0.965 (external)
Park J. et al. (2025) [18]	H&E WSI	Multi-country, multi-scanner	1,461 (+599 external)	Lunit SCOPE Genotype Predictor (DL)	0.905 (primary); 0.860 (external)
Wang S. et al. (2019) [19]	CT	Two hospitals (China)	844 patients / 14,926 images	End-to-end deep CNN	0.85 (primary); 0.81 (validation)
Zhang B. et al. (2021) [20]	CT radiomics	Two hospitals (China)	709 + 205 patients	SE-CNN on radiomics feature maps	0.910 (internal); 0.841 (external)
Jiang X. et al. (2025) [21]	CT radiomics + DL	Single-center	220 patients	ViT + radiomics early fusion	0.910 (fusion model)
Huang L. et al. (2025) [22]	CT radiomics + DL	Two hospitals	826 patients	3D CNN + radiomics, soft-voting fusion	0.925 (internal); 0.889 (external)
Yoon H. et al. (2022) [23]	CT (pure GGN)	Single-center	185 resected pGGNs	CNN + clinical factors	0.85 (test set)

DEEP LEARNING FOR EGFR MUTATION PREDICTION IN LUNG ADENOCARCINOMA

Study (Year)	Modality	Data Source / Cohort	Sample Size	Model / Architecture	Best Reported AUC
Liu Q. et al. (2020) [24]	18F-FDG PET/CT	Two hospitals	148 patients	Radiomics + XGBoost	0.87 (overall)
Wang Y. et al. (2025) [25]	18F-FDG PET/CT	Single-center	218 patients	ConvNeXt DL + radiomics + clinical fusion	0.821 (CRD model)
Cao P. et al. (2025) [26]	MRI (brain metastasis)	Single-center	288 patients	ResNet34 features + LASSO radiomics	0.936 (validation)
Wu J. et al. (2024) [27]	CT radiomics + DL	Four centers	438 patients	Habitat radiomics + DL nomogram	0.917 (training); 0.809 (external)
Yao X. et al. (2024) [28]	18F-FDG PET/CT	Single-center	202 patients	ResNet deep features + PyRadiomics + SVM	0.94 (PET/CT fusion)
Shao X. et al. (2024) [29]	18F-FDG PET/CT + clinical	Two centers	516 patients	3D CNN (transfer learning) + clinical fusion	0.883 (training); 0.730 (test)
Xu J. et al. (2025) [30]	CT + WSI (multimodal)	Two institutions	396 patients	Deep learning multimodal feature-interaction fusion	0.856 (internal); 0.817 (external)

† Conference abstract with an extractable metric, included per the eligibility criteria in Section 2.3; no accompanying full-length peer-reviewed report was identified for this study.

Reported AUC also varied by modality. Histopathology-based models reached a median AUC of 0.826 (mean 0.828; range 0.682-0.933) across seven studies. Radiology-based models reached a median AUC of 0.910 (mean 0.892; range 0.821-0.940) across eleven studies. Across all nineteen studies with a comparable AUC, the median was 0.883 (mean 0.867; range 0.682-0.940). Radiology-based models reached a somewhat higher median AUC than histopathology-based models, though the two distributions overlap substantially, as Figure 3 shows.

Figure 3 compares the reported AUC point estimates across all nineteen studies, grouped by modality. Confidence intervals are not shown, because most studies did not report per-study variance. A formal pooled meta-analysis, with 95% confidence intervals and I^2 heterogeneity, was not performed for this reason. Two related meta-analyses already cover parts of this literature. Ma et al. pooled seventeen 18F-FDG PET/CT radiomics studies and reported a sensitivity of 0.76, specificity of 0.78, and AUC of 0.84 for EGFR prediction [8]. Fuster-Matanzo et al. pooled fifty-one CT radiomics studies across EGFR, ALK, and KRAS, and reported an EGFR sensitivity of 0.754 for radiomics-only models, rising to 0.806 once clinical data were added [10].

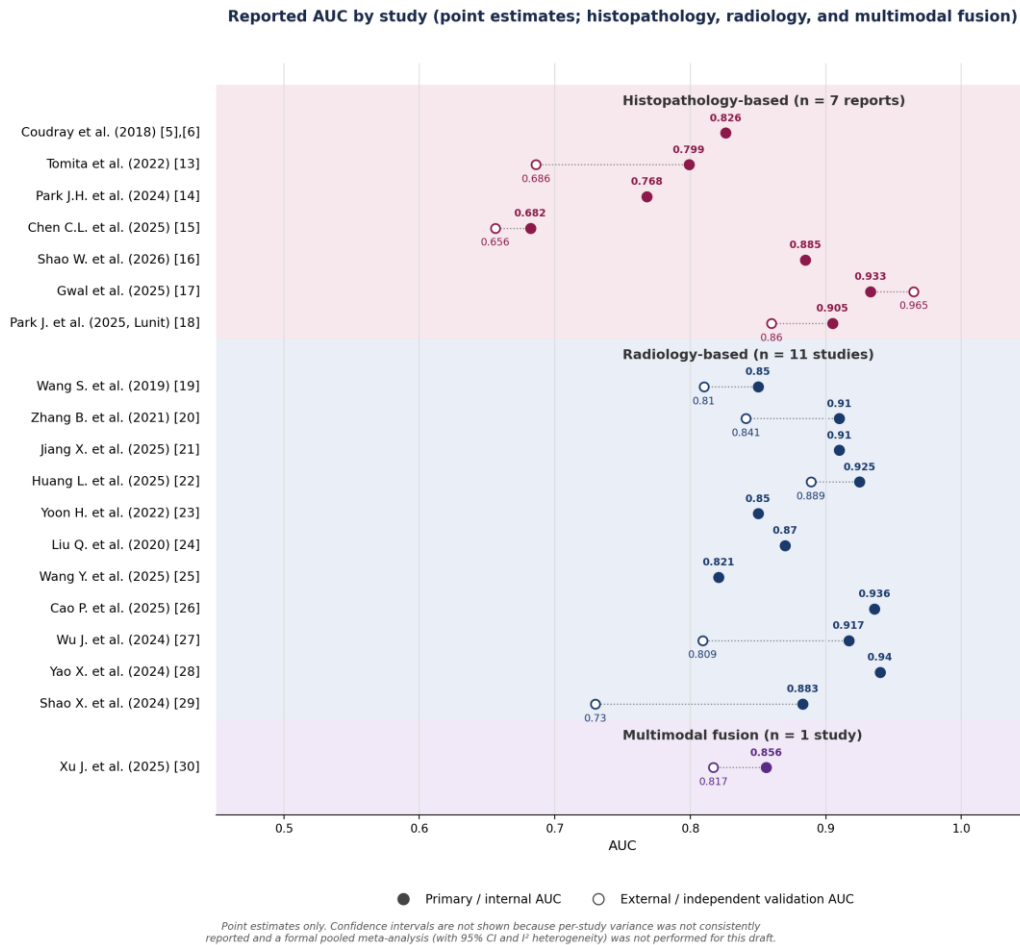


Figure 3. Reported AUC point estimates for the nineteen included studies, grouped by modality. Filled circles show the primary or internal AUC; hollow circles show external or independent validation AUC, where reported. This is a descriptive comparison, not a pooled meta-analytic estimate.

3.2 Histopathology-Based EGFR Prediction

Coudray et al. trained an Inception-v3 CNN on TCGA whole-slide images. The model classified tumor subtype and predicted the ten most commonly mutated genes in LUAD. For EGFR, it reached an AUC of 0.826-0.856 on held-out test data [5],[6].

Later studies used attention-based multiple instance learning (MIL). MIL aggregates thousands of image patches per slide into one slide-level prediction, and also assigns attention weights to informative regions. Tomita et al. reached an AUC of 0.799 on an internal cohort, but only 0.686 on external CPTAC-3 validation [13]. This gap between internal and external performance repeats across the literature.

Park et al. built an EGFR mutation prevalence (EMP) score from patch-level MIL predictions.

High EMP scores matched papillary and acinar histology patterns. Low scores matched solid growth patterns [14]. This gave a biologically plausible basis for the model's predictions.

Recent studies use stronger backbones, such as DenseNet201 or Vision Transformer-based foundation model encoders, on larger or multi-center training sets. Shao et al. reached an AUC of 0.885 using attention-based MIL with DenseNet201 encoding and logistic-regression aggregation [16]. Gwal et al. used a pathology foundation-model encoder with attention-based MIL on a small Indian cohort of 170 slides. The model reached an AUC of 0.933 internally and 0.965 on an independent TCGA external test set [17]. Park (Lunit SCOPE Genotype Predictor) trained on more than 12,000 WSIs and validated across 1,461 and 599 patients from eleven countries. The model reached AUCs of 0.905 and 0.860, with 90.4% scanner concordance [18].

In contrast, weakly supervised pipelines without annotation reached lower AUCs of 0.66-0.68 for EGFR prediction on smaller single-center cohorts. The same pipelines reached an AUC above 0.95 for the easier task of subtyping adenocarcinoma versus squamous cell carcinoma [15]. This shows that EGFR prediction is harder than subtyping.

Ho et al. used deep segmentation of growth patterns. Acinar patterns were enriched for EGFR mutations. Solid patterns were depleted of EGFR mutations and enriched for TP53 mutations instead [31]. This gives a pattern-level explanation for why EGFR status is learnable from tissue shape.

3.3 Radiology-Based EGFR Prediction

Wang et al. trained an end-to-end CNN on 14,926 CT images from 844 patients. The model reached AUCs of 0.85 and 0.81 in primary and independent validation cohorts [19]. The deep-learning score differed between EGFR-mutant and wild-type tumors.

Radiomics-only models extract hand-crafted features, such as texture and shape, from segmented tumors. These models generally perform worse than deep-learning or fusion models. Zhang et al. found that a CNN using a radiomics feature map (AUC 0.910 internal, 0.841 external) outperformed a plain CNN and classical machine-learning baselines [20].

Fusion of radiomics and deep-learning features gives consistently better results than either alone.

Jiang et al. found that a Vision Transformer fused with radiomics reached an AUC of 0.910, higher than radiomics alone (0.792) or the Vision Transformer alone (0.870) [21]. Huang et al. tested intratumoral and peritumoral regions across five margin widths. A soft-voting ensemble of 3D deep-learning and radiomics models reached AUCs of 0.925 internal and 0.889 external [22].

PET/CT studies show the same pattern. Wang et al. found that adding a deep ConvNeXt-derived score to a clinical-radiomics model raised the AUC from 0.739 to 0.821 [25]. A meta-analysis of seventeen 18F-FDG PET/CT radiomics studies (3,763 patients) reported a pooled sensitivity of 0.76, specificity of 0.78, and AUC of 0.84 for EGFR prediction [8]. Deep-learning-based feature extraction was a major source of heterogeneity across these studies.

Some studies focused on CT phenotypes linked to EGFR status. For pure ground-glass opacity nodules, deep learning combined with clinical variables reached an AUC of 0.85. This is higher than radiomics alone (0.64) or clinical variables alone (0.50) [23]. For brain metastasis from lung adenocarcinoma, a deep-learning-augmented MRI radiomics signature reached a validation AUC of 0.936 [26].

A broader systematic review and meta-analysis covered 124 studies (51 meta-analyzed) predicting EGFR, ALK, and KRAS status from CT radiomics in NSCLC. For EGFR, the pooled sensitivity was 0.754 for radiomics-only models, rising to 0.806 when clinical data were added [10]. Meta-regression found no significant benefit from adding clinical variables. This differs somewhat from the fusion-model gains seen in the individual studies above, but supports the same general direction: models that combine feature types outperform single-feature models.

Three further studies extend this fusion pattern. Wu et al. combined habitat radiomics, deep learning, and clinical variables in a four-center study of 438 patients. Their nomogram reached AUCs of 0.917, 0.837, and 0.809 in training, validation, and external test sets [27]. Yao et al. fused shallow radiomics features with deep ResNet features from 18F-FDG PET/CT in 202 patients. The combined model reached an AUC of 0.94, well above shallow features alone (0.68) [28]. Shao et al. built a three-stream model fusing PET, CT, and clinical data with transfer learning in 516 patients, reaching an AUC of 0.883 in training but only 0.730 on an independent test set [29]. This

test-set drop again illustrates the internal-to-external generalization gap seen throughout this literature.

3.4 Multimodal (Radiology + Histopathology) Fusion

One study in this review combined radiology and histopathology directly, rather than treating them as separate modalities. Xu et al. built a deep learning multimodal feature-interaction fusion (DL-MFIF) model. It integrates macroscopic CT features with microscopic whole-slide histopathology features, in 396 patients with advanced-stage lung adenocarcinoma from two institutions [30]. The fused model reached an AUC of 0.856 internally and 0.817 on external validation. This outperformed single-modality baselines built from the same data.

This is an early proof of concept, not an established approach. The cohort was restricted to advanced-stage disease, which raises concern about spectrum bias. Only one study in this review attempted this kind of fusion. Combining radiology and histopathology in one model remains a promising but largely untested direction for EGFR prediction.

3.5 Model Architectures and Methodological Trends

The histopathology literature shows three model generations. The first used patch-level CNNs, such as Inception-v3, with simple average pooling to slide level. The second used attention-based MIL frameworks. Clustering-constrained-attention MIL (CLAM) is the most influential example. CLAM uses attention to classify slides and locate relevant regions, using only slide-level labels [32]. Later work refined instance discrimination, attribute-aware attention, and prototype-based interpretability [33], [34].

One large benchmarking study compared MIL and classical weakly supervised pipelines across six mutation-prediction tasks. For mutation prediction specifically, classical pipelines outperformed several MIL methods, despite MIL's theoretical advantages [35]. This shows that empirical benchmarking matters more than architectural complexity alone.

The third generation uses large self-supervised pathology foundation models as feature extractors. Examples include UNI, pretrained on over 100 million images from 100,000 WSIs [36]; CONCH, a vision-language model pretrained on 1.17 million image-caption pairs [37]; CHIEF, pretrained

on 60,530 WSIs across nineteen anatomical sites [38]; and TITAN, a multimodal model that adds pathology reports and synthetic captions [39]. The two studies in this review that used strong pretrained encoders, Gwal et al. and Shao W. et al., rank among the strongest histopathology results rather than training a CNN from scratch [17], [16]. This matches independent benchmarking showing that foundation-model feature extractors outperform end-to-end training in low-data clinical tasks [40], [41].

3.6 Interpretability and Model Transparency

Most included studies addressed interpretability through attention heatmaps or Grad-CAM saliency maps. Attention heatmaps were used to check that high-attention regions matched known histology patterns, such as papillary or acinar architecture in EGFR-mutant tumors [14], [31]. Shao W. et al. also reported calibration curves and decision curve analysis to assess clinical benefit beyond AUC alone [16].

The wider explainable-AI literature raises some caution. Comparative studies found that Grad-CAM saliency maps often lack the detail needed for confident clinical interpretation, compared with pixel-level methods [42]. Grad-CAM's explanatory value in medical imaging remains contested [43]. A meta-analysis of 67 studies across radiology, pathology, and ophthalmology found that common post-hoc methods (SHAP, LIME, Grad-CAM) vary in fidelity and stability under noise. Inherently interpretable models typically lose 5-7% AUC compared to black-box models [44]. None of the EGFR-specific studies in this review tested the fidelity or stability of their explanations. This is a clear gap between the wider XAI literature and current practice in this application.

4. DISCUSSION

This review addressed three questions, stated in Section 1. For RQ1, histopathology-based models used CNNs, attention-based MIL, and pathology foundation models, with AUCs from 0.682 to 0.933. Radiology-based models used radiomics, deep features, and their fusion, with AUCs from 0.821 to 0.940. Fusion approaches and foundation-model encoders gave the strongest and most

externally stable results. For RQ2, most studies used attention heatmaps or Grad-CAM for interpretability. Few tested whether these explanations were reliable, and few applied a formal risk-of-bias tool. For RQ3, the main barriers are small, single-institution cohorts, limited external validation, and untested interpretability methods. The rest of this section discusses these findings in more depth.

Table 4 synthesizes the key differences across the three modalities covered in this review.

Table 4. Cross-modality synthesis of histopathology-, radiology-, and multimodal-based EGFR prediction.

Dimension	Histopathology	Radiology	Multimodal
Studies (n)	7	11	1
Median AUC (mean; range)	0.826 (0.828; 0.682-0.933)	0.910 (0.892; 0.821-0.940)	0.856 (single study)
Risk of bias (QUADAS-2)	1 low, 6 some concerns	9 some concerns, 2 high	1 high
Dominant architecture trend	CNN, then attention-based MIL, then pathology foundation-model encoders	Hand-crafted radiomics, then radiomics fused with deep-learning features	Feature-interaction fusion combining CT and whole-slide histopathology
Foundation-model use	2 of 7 studies (Gwal et al.; Shao W. et al.)	None reported	None reported
Key insight	Foundation-model encoders gave the strongest and most externally stable results, despite small training sets.	Fusing radiomics with deep-learning features consistently outperformed either feature type alone.	An early proof of concept; only one restricted, advanced-stage cohort has been tested.

This review combined nineteen studies on deep learning for EGFR prediction from lung adenocarcinoma images. Three patterns stand out. First, performance is consistently above chance and often clinically useful (AUC 0.682-0.940). Performance depends strongly on cohort size, imaging modality, and whether external, multi-institution validation was done. Studies validated only on internal splits from one institution reported AUCs 5-15 points higher than the same or similar models tested on external cohorts [13,20,27,29]. This gap likely reflects institution-specific staining, scanner hardware, and patient population, rather than a true difference in how learnable EGFR-linked morphology is.

Second, fusion consistently outperforms single-modality approaches. This holds for radiomics fused with deep features in radiology [21], [22], [25], [27], [28], and for strong pretrained encoders fused with attention-based aggregation in histopathology [16], [17]. One study went a step further and fused radiology with histopathology directly, reaching an AUC of 0.86 [30]. This is only one

study, on a restricted advanced-stage cohort, so it should be read as an early signal rather than a proven approach. Hand-crafted and learned features appear to carry complementary information across modalities. Future work should extend this cross-modality fusion to larger, more representative cohorts.

Third, the field is moving from bespoke CNN training toward transfer learning from large pathology foundation models. This shift matters for EGFR prediction, a domain-specific, rare-outcome task. Most single institutions have only a few hundred mutation-labeled slides, too few to train a large network from scratch without overfitting. Foundation-model-based approaches in this review reached among the strongest and most stable results, despite small task-specific training sets [17]. This matches broader benchmarking evidence that pretraining data diversity, not task-specific data volume, drives downstream performance in computational pathology [40].

Population representativeness is a separate concern for EGFR prediction. EGFR mutation prevalence differs between Asian and non-Asian populations [3,4]. A model trained mostly on one population's tissue and imaging features may not transfer to a population with different mutation prevalence and tumor biology. Only one study in this review tested cross-population generalization, using an Indian cohort validated against the mostly Western TCGA dataset [17]. This remains an underexplored area, especially for research groups in Southeast Asia and other under-represented settings.

Interpretability practice in this literature lags behind the wider explainable-AI field. Attention heatmaps and Grad-CAM visualizations are common, but no EGFR-prediction study tested their fidelity or stability. The wider medical-imaging XAI literature shows that such explanations can be unstable or only weakly informative about model reasoning [42], [43], [44]. Clinical trust and regulatory approval will likely depend on faithful explanations, not just visually plausible ones. This is a clear and fixable gap for future EGFR studies.

5. LIMITATIONS

This review used one search platform: Consensus, which combines Semantic Scholar, PubMed,

Scopus, and arXiv. It did not run separate, reproducible searches in each database with exact, date-stamped queries. This fits the scope of a rapid review, but it limits reproducibility. A full systematic review should redo the search per database. It should document the exact search strings, dates, and hit counts, with a full PRISMA flow diagram.

Figure 3 shows AUC point estimates only. Most included studies did not report a confidence interval or standard error for their AUC. Without this variance data, a formal pooled meta-analysis with 95% confidence intervals and I^2 heterogeneity could not be computed. A future update should contact study authors or extract raw confusion-matrix data where possible. This would allow proper pooling.

The QUADAS-2 risk-of-bias assessment (Table 4) was performed by a single reviewer rather than independently by two reviewers with adjudication of disagreements, which is the recommended practice for systematic reviews. This may introduce assessor-specific bias into the risk-of-bias ratings reported here.

6. CONCLUSION

This review analyzed nineteen studies on deep learning for EGFR mutation prediction from histopathology and radiology images in lung adenocarcinoma. Attention-based multiple instance learning, pathology foundation models, and radiomics-deep learning fusion produced the strongest and most externally valid results, with AUCs up to 0.933-0.940 in the best-validated studies. A single early study also showed that fusing radiology with histopathology directly is feasible, reaching an AUC of 0.86. Small cohorts, limited external validation, spectrum bias in some subgroup-restricted studies, and untested interpretability methods remain the main barriers to clinical use.

Future work should prioritize four directions. First, multi-institution cohorts with pre-registered external validation. Second, systematic benchmarking of pathology and radiology foundation models specifically for EGFR prediction. Third, larger-scale radiology-histopathology fusion studies, built on more representative, non-restricted cohorts. Fourth, quantitative testing of

attention- and Grad-CAM-based explanations, rather than visual inspection alone. Addressing these gaps would move image-based EGFR prediction from a promising research direction toward a validated adjunct to molecular testing.

ACKNOWLEDGEMENT

This research was funded by the Ministry of Higher Education, Research and Technology of the Republic of Indonesia through the Fundamental Research Grant under the BOPTN scheme, Project Number B/036/UN46.1/PT.01.03/BIMA/PL-Multitahunlanjutan/2026.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE MANUSCRIPT

During the preparation of this work, the authors used Claude (Anthropic) to assist with drafting and editing text, condensing and restructuring manuscript sections, and checking consistency in terminology and structure across sections. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

- [1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R.L. Siegel, et al., Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, *CA Cancer J. Clin.* 74 (2024), 229-263. <https://doi.org/10.3322/caac.21834>.
- [2] A.C.Z. Gelatti, A. Drilon, F.C. Santini, Optimizing the Sequencing of Tyrosine Kinase Inhibitors (TKIs) in Epidermal Growth Factor Receptor (EGFR) Mutation-Positive Non-Small Cell Lung Cancer (NSCLC), *Lung Cancer* 137 (2019), 113-122. <https://doi.org/10.1016/j.lungcan.2019.09.017>.
- [3] Y. Shi, J. Li, S. Zhang, M. Wang, S. Yang, et al., Molecular Epidemiology of EGFR Mutations in Asian Patients with Advanced Non-Small-Cell Lung Cancer of Adenocarcinoma Histology – Mainland China Subset Analysis of the PIONEER Study, *PLoS One* 10 (2015), e0143515. <https://doi.org/10.1371/journal.pone.0143515>.
- [4] F.C.A. de Moraes, A.L.S. de Oliveira Rodrigues, E. Pasqualotto, J.F. Cassemiro, J.W.L. Choque, et al., Ethnic

DEEP LEARNING FOR EGFR MUTATION PREDICTION IN LUNG ADENOCARCINOMA

- Disparities in Survival and Progression Among EGFR-Mutated Adenocarcinoma of Lung Cancer Patients Treated with Tyrosine Kinase Inhibitors: A Systematic Review and Meta-Analysis, *Clin. Transl. Oncol.* 27 (2025), 3001-3014. <https://doi.org/10.1007/s12094-024-03843-4>.
- [5] N. Coudray, P.S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, et al., Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images Using Deep Learning, *Nat. Med.* 24 (2018), 1559-1567. <https://doi.org/10.1038/s41591-018-0177-5>.
- [6] N. Coudray, A.L. Moreira, T. Sakellaropoulos, D. Fenyő, N. Razavian, et al., Determining EGFR and STK11 Mutational Status in Lung Adenocarcinoma Histopathology Images Using Deep Learning, *Cancer Res.* 78 (2018), 5309. <https://doi.org/10.1158/1538-7445.AM2018-5309>.
- [7] X. Ge, J. Gao, R. Niu, Y. Shi, X. Shao, et al., New Research Progress on 18F-FDG PET/CT Radiomics for EGFR Mutation Prediction in Lung Adenocarcinoma: a Review, *Front. Oncol.* 13 (2023), 1242392. <https://doi.org/10.3389/fonc.2023.1242392>.
- [8] N. Ma, W. Yang, Q. Wang, C. Cui, Y. Hu, et al., Predictive Value of 18F-FDG PET/CT Radiomics for EGFR Mutation Status in Non-Small Cell Lung Cancer: a Systematic Review and Meta-Analysis, *Front. Oncol.* 14 (2024), 1281572. <https://doi.org/10.3389/fonc.2024.1281572>.
- [9] H.S. Nguyen, D.K.N. Ho, N.N. Nguyen, H.M. Tran, K.-W. Tam, et al., Predicting EGFR Mutation Status in Non-Small Cell Lung Cancer Using Artificial Intelligence: A Systematic Review and Meta-Analysis, *Acad. Radiol.* 31 (2024), 660-683. <https://doi.org/10.1016/j.acra.2023.03.040>.
- [10] A. Fuster-Matanzo, A. Picó-Peris, F. Bellvís-Bataller, A. Jimenez-Pastor, G.J. Weiss, et al., Prediction of Oncogene Mutation Status in Non-Small Cell Lung Cancer: a Systematic Review and Meta-Analysis with a Special Focus on Artificial Intelligence-Based Methods, *Eur. Radiol.* 36 (2026), 2157-2185. <https://doi.org/10.1007/s00330-025-11962-x>.
- [11] R. Parra-Medina, G. Guerron-Gomez, D. Mendivelso-González, J.H. Gil-Gómez, J.P. Alzate, et al., Deep Learning in Histopathology Images for Prediction of Oncogenic Driver Molecular Alterations in Lung Cancer: a Systematic Review and Meta-Analysis, *Transl. Lung Cancer Res.* 14 (2025), 1756-1769. <https://doi.org/10.21037/tlcr-2024-1196>.
- [12] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, et al., The PRISMA 2020 Statement: An

Updated Guideline for Reporting Systematic Reviews, *PLoS Med.* 18 (2021), e1003583.

<https://doi.org/10.1371/journal.pmed.1003583>.

- [13] N. Tomita, L.J. Tafe, A.A. Suriawinata, G.J. Tsongalis, M. Nasir-Moin, et al., Predicting Oncogene Mutations of Lung Cancer Using Deep Learning and Histopathologic Features on Whole-Slide Images, *Transl. Oncol.* 24 (2022), 101494. <https://doi.org/10.1016/j.tranon.2022.101494>.
- [14] J.H. Park, J.H. Lim, S. Kim, C.-H. Kim, J.-S. Choi, et al., Deep Learning-Based Analysis of EGFR Mutation Prevalence in Lung Adenocarcinoma H&E Whole Slide Images, *J. Pathol. Clin. Res.* 10 (2024), e70004. <https://doi.org/10.1002/2056-4538.70004>.
- [15] C.-L. Chen, C.-C. Chen, T.-H. Chang, M.-S. Hsieh, C.-Y. Yeh, et al., Weakly Supervised Prediction of EGFR Mutation Status and Early Metastasis of Lung Adenocarcinoma Whole Slide Histopathology Images Using Artificial Intelligence, *Clin. Cancer Res.* 31 (2025), A048. <https://doi.org/10.1158/1557-3265.AIMACHINE-A048>.
- [16] W. Shao, W. Gu, S. Song, D. Chen, Z. Shao, et al., Deep Learning Framework for Predicting EGFR Mutation Status from H&E Whole Slide Images in Lung Adenocarcinoma, *BMC Cancer* 26 (2026), 601. <https://doi.org/10.1186/s12885-026-15957-9>.
- [17] S.S. Gwal, Rajan, S. Devgan, S. Satapathy, A. Goyal, et al., Predicting EGFR Mutation in LUAD from Histopathological Whole-Slide Images Using Pretrained Foundation Model and Transfer Learning: An Indian Cohort Study, *arXiv:2508.01352*, 2025. <https://doi.org/10.48550/arXiv.2508.01352>.
- [18] J. Park, S. Shin, W. Hwang, S. Keum, B. Brattoli, et al., Deep Learning Predicts EGFR Mutation Status from Histology Images in Non-Small Cell Lung Cancer, *Cancer Res. Commun.* 5 (2025), 2127-2141. <https://doi.org/10.1158/2767-9764.CRC-25-0155>.
- [19] S. Wang, J. Shi, Z. Ye, D. Dong, D. Yu, et al., Predicting EGFR Mutation Status in Lung Adenocarcinoma on Computed Tomography Image Using Deep Learning, *Eur. Respir. J.* 53 (2019), 1800986. <https://doi.org/10.1183/13993003.00986-2018>.
- [20] B. Zhang, S. Qi, X. Pan, C. Li, Y. Yao, et al., Deep CNN Model Using CT Radiomics Feature Mapping Recognizes EGFR Gene Mutation Status of Lung Adenocarcinoma, *Front. Oncol.* 10 (2021), 598721. <https://doi.org/10.3389/fonc.2020.598721>.

- [21] X. Jiang, Q. Sun, C. Wang, W. Li, W. Chen, et al., CT-Based Radiomics and Deep Learning to Predict EGFR Mutation Status in Lung Adenocarcinoma, *Front. Oncol.* 15 (2025), 1597548.
<https://doi.org/10.3389/fonc.2025.1597548>.
- [22] L. Huang, L. Xu, X. Wang, G. Zhang, X. Gao, et al., Prediction of EGFR Mutations in Lung Adenocarcinoma via CT Images: A Comparative Study of Intratumoral and Peritumoral Radiomics, Deep Learning, and Fusion Models, *Acad. Radiol.* 32 (2025), 4880-4892. <https://doi.org/10.1016/j.acra.2025.04.029>.
- [23] H.J. Yoon, J. Choi, E. Kim, S.-W. Um, N. Kang, et al., Deep Learning Analysis to Predict EGFR Mutation Status in Lung Adenocarcinoma Manifesting as Pure Ground-Glass Opacity Nodules on CT, *Front. Oncol.* 12 (2022), 951575. <https://doi.org/10.3389/fonc.2022.951575>.
- [24] Q. Liu, D. Sun, N. Li, J. Kim, D. Feng, et al., Predicting EGFR Mutation Subtypes in Lung Adenocarcinoma Using 18F-FDG PET/CT Radiomic Features, *Transl. Lung Cancer Res.* 9 (2020), 549-562.
<https://doi.org/10.21037/tlcr.2020.04.17>.
- [25] Y. Wang, Z. Chen, J. Li, Y. Cai, C. Sun, et al., An Integrated Clinical-Radiomics-Deep Learning Model Based on 18F-FDG PET/CT for Predicting EGFR Mutation Status in Lung Adenocarcinoma, *Cancer Med.* 14 (2025), e71370. <https://doi.org/10.1002/cam4.71370>.
- [26] P. Cao, X. Jia, X. Wang, L. Fan, Z. Chen, et al., Deep Learning Radiomics for the Prediction of Epidermal Growth Factor Receptor Mutation Status Based on MRI in Brain Metastasis from Lung Adenocarcinoma Patients, *BMC Cancer* 25 (2025), 443. <https://doi.org/10.1186/s12885-025-13823-8>.
- [27] J. Wu, H. Meng, L. Zhou, M. Wang, S. Jin, et al., Habitat Radiomics and Deep Learning Fusion Nomogram to Predict EGFR Mutation Status in Stage I Non-Small Cell Lung Cancer: A Multicenter Study, *Sci. Rep.* 14 (2024), 15877. <https://doi.org/10.1038/s41598-024-66751-1>.
- [28] X. Yao, Y. Zhu, Z. Huang, Y. Wang, S. Cong, et al., Fusion of Shallow and Deep Features from 18F-FDG PET/CT for Predicting EGFR-Sensitizing Mutations in Non-Small Cell Lung Cancer, *Quant. Imaging Med. Surg.* 14 (2024), 5460-5472. <https://doi.org/10.21037/qims-23-1028>.
- [29] X. Shao, X. Ge, J. Gao, R. Niu, Y. Shi, et al., Transfer Learning-Based PET/CT Three-Dimensional Convolutional Neural Network Fusion of Image and Clinical Information for Prediction of EGFR Mutation in Lung Adenocarcinoma, *BMC Med. Imaging* 24 (2024), 54. <https://doi.org/10.1186/s12880-024-01232-5>.

- [30] J. Xu, B. Feng, X. Chen, F. Wu, Y. Liu, et al., Deep Learning-Based Multimodal Feature Interaction-Guided Fusion: Enhancing the Evaluation of EGFR in Advanced Lung Adenocarcinoma, *Acad. Radiol.* 32 (2025), 5585-5595. <https://doi.org/10.1016/j.acra.2025.04.071>.
- [31] D.J. Ho, J.C. Chang, R.G. Aly, H.C.T. Nguyen, P.S. Adusumilli, et al., Deep Learning–Based Segmentation of Lung Adenocarcinoma Whole-Slide Images for Objective Grading, Tumor Spread Through Air Spaces Identification, and Mutation Prediction, *Mod. Pathol.* 38 (2025), 100907. <https://doi.org/10.1016/j.modpat.2025.100907>.
- [32] M.Y. Lu, D.F.K. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, et al., Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images, *Nat. Biomed. Eng.* 5 (2021), 555-570. <https://doi.org/10.1038/s41551-020-00682-w>.
- [33] L. Cai, S. Huang, Y. Zhang, J. Lu, Y. Zhang, AttriMIL: Revisiting Attention-Based Multiple Instance Learning for Whole-Slide Pathological Image Classification from a Perspective of Instance Attributes, *Med. Image Anal.* 103 (2025), 103631. <https://doi.org/10.1016/j.media.2025.103631>.
- [34] J. Liu, A. Mao, Y. Niu, X. Zhang, T. Gong, et al., PAMIL: Prototype Attention-Based Multiple Instance Learning for Whole Slide Image Classification, in: M.G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, et al., (eds) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, pp. 362-372, (2024). https://doi.org/10.1007/978-3-031-72083-3_34.
- [35] N. Ghaffari Laleh, H.S. Muti, C.M.L. Loeffler, A. Echle, O.L. Saldanha, et al., Benchmarking Weakly-Supervised Deep Learning Pipelines for Whole Slide Classification in Computational Pathology, *Med. Image Anal.* 79 (2022), 102474. <https://doi.org/10.1016/j.media.2022.102474>.
- [36] R.J. Chen, T. Ding, M.Y. Lu, D.F.K. Williamson, G. Jaume, et al., Towards a General-Purpose Foundation Model for Computational Pathology, *Nat. Med.* 30 (2024), 850-862. <https://doi.org/10.1038/s41591-024-02857-3>.
- [37] M.Y. Lu, B. Chen, D.F.K. Williamson, R.J. Chen, I. Liang, et al., A Visual-Language Foundation Model for Computational Pathology, *Nat. Med.* 30 (2024), 863-874. <https://doi.org/10.1038/s41591-024-02856-4>.
- [38] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, et al., A Pathology Foundation Model for Cancer Diagnosis

- and Prognosis Prediction, *Nature* 634 (2024), 970-978. <https://doi.org/10.1038/s41586-024-07894-z>.
- [39] T. Ding, S.J. Wagner, A.H. Song, R.J. Chen, M.Y. Lu, et al., A Multimodal Whole-Slide Foundation Model for Pathology, *Nat. Med.* 31 (2025), 3749-3761. <https://doi.org/10.1038/s41591-025-03982-3>.
- [40] P. Neidlinger, O.S.M. El Nahhas, H.S. Muti, T. Lenz, M. Hoffmeister, et al., Benchmarking Foundation Models as Feature Extractors for Weakly Supervised Computational Pathology, *Nat. Biomed. Eng.* 10 (2026), 1113-1123. <https://doi.org/10.1038/s41551-025-01516-3>.
- [41] G. Campanella, S. Chen, M. Singh, R. Verma, S. Muehlstedt, et al., A Clinical Benchmark of Public Self-Supervised Pathology Foundation Models, *Nat. Commun.* 16 (2025), 3640. <https://doi.org/10.1038/s41467-025-58796-1>.
- [42] M. Ennab, H. Mcheick, Advancing AI Interpretability in Medical Imaging: a Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models, *Mach. Learn. Knowl. Extr.* 7 (2025), 12. <https://doi.org/10.3390/make7010012>.
- [43] S. Suara, A. Jha, P. Sinha, A.A. Sekh, Is Grad-CAM Explainable in Medical Images?, *arXiv:2307.10506*, 2023. <https://doi.org/10.48550/arXiv.2307.10506>.
- [44] Y. Singh, Q.A. Hathaway, V. Keishing, S. Salehi, Y. Wei, et al., Beyond Post hoc Explanations: a Comprehensive Framework for Accountable AI in Medical Imaging Through Transparency, Interpretability, and Explainability, *Bioengineering* 12 (2025), 879. <https://doi.org/10.3390/bioengineering12080879>.