DEMO: A Scalable Artificial Intelligence Framework for Rapid EGFR Mutation Screening in Lung Cancer

Neeraj Kumar[1,*] **Swaraj Nanda**[1] **Siddharth Singi**[1] **Gabriele Campanella**[2]

Thomas Fuchs[2] **Chad Vanderbilt**[1,*]

Abstract

This paper presents two approaches to predict Epidermal Growth Factor Receptor (EGFR) mutation status in non-small cell lung cancer (NSCLC) patients from Hematoxylin and Eosin (H&E) stained histopathology whole slide images (WSIs). The first uses a two-step process: training a vision transformer on histology classification, then using it as a frozen feature extractor for a multiple instance learning (MIL) aggregator. The second implements end-to-end training of a pre-trained foundation model encoder and an MIL aggregator using distributed training. An in-real-time pipeline is presented for rapid clinical EGFR screening. Experiments on a large patient cohort demonstrate effectiveness, with the best model achieving 0.83 AUC and 2-minute inference time per slide, offering a potential rapid, cost-effective alternative to conventional molecular testing in a live clinical setting.

1 Introduction

Lung cancer is the deadliest cancer in the United States (US). EGFR testing is the cornerstone of early treatment protocols and determine first line of therapy with EGFR tyrosine kinase inhibitors (TKIs) [Network, 2024]. This targeted approach leads to improved progression-free survival, enhanced quality of life, and potentially reduced overall disease burden in patients with EGFR-mutant NSCLC [Riely et al., 2024, Ramalingam et al., 2020]. Only 22% of advanced NSCLC patients in the US had access to NGS testing between 2010-2018, with even lower rates in other developed countries [Schilsky and Longo, 2022, Normanno et al., 2022]. Limited adoption stems from high costs, infrastructure gaps, tissue sampling issues, and time constraints [Hiley et al., 2016]. Where NGS is unavailable, less sensitive rapid tests are used [Grant et al., 2022, Momeni-Boroujeni et al., 2021].

Given these constraints, there is a need for complementary techniques to efficiently screen NSCLC patients for EGFR status prediction. Artificial intelligence (AI)-based approaches utilizing digital whole slide images (WSIs) of routine hematoxylin and eosin (H&E) stained histopathology slides have emerged as a promising solution [Coudray et al., 2018]. AI-based EGFR screening requires only digitized H&E slides, making it potentially cost-effective and accessible even in resource-limited settings. This could serve as a valuable adjunct to genomic testing allowing for lower cost, faster turnaround time, and better tissue utilization. Based on these considerations, this paper presents: (a) Development and validation of two AI-based approaches to predict EGFR mutation status directly from WSIs of NSCLC patients. (b) Comprehensive evaluation of the proposed EGFR prediction

* Corresponding authors: kumarn6@mskcc.org (N.K.), vanderbc@mskcc.org (C.V.)
[1] Memorial Sloan Kettering Cancer Center, New York, NY 10065
[2] Icahn School of Medicine at Mount Sinai, New York, NY 10029

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

methods utilizing one of the largest retrospective patient cohorts. (c) A demonstration of the clinical implementation of an EGFR status prediction system, designed to assess WSIs of NSCLC patients in real-time and at scale.

2 Background

Recent advances in deep learning have shown great promise for automated analysis of histopathology WSIs. However, WSIs are gigapixel images, often containing around 100,000 pixels per dimension at $20\times$ magnification (0.5 microns per pixel), making pixel-level annotations time-consuming and expensive. To overcome this, researchers have developed weakly supervised learning techniques that rely on slide-level labels instead of pixel-level annotations.

2.1 Weakly Supervised Learning

Due to the large size of WSIs, slides are typically divided into tens of thousands of smaller patches, which are then analyzed using multiple instance learning (MIL). MIL formulates WSI classification as a bag-of-instances problem, where each WSI is a bag containing many patches (instances). A WSI is considered positive if at least one patch is positive. Ilse et al. proposed an attention-based MIL approach that learns to identify the most informative regions in a WSI for breast and colon cancer detection [Ilse et al., 2018]. Building on this work, Campanella et al. developed a deep learning system trained on over 44,000 WSIs using only slide-level labels, achieving performance comparable to pathologists in cancer diagnosis [Campanella et al., 2019].

Most recently, various foundation models developed through self-supervised learning have been shown to be powerful feature extractors, delivering better performance in MIL-based methods for various downstream pathology tasks. Virchow, UNI, and Prov-GigaPath are vision transformer models pretrained on large datasets of histopathology slides using self-supervised learning techniques. Virchow, trained on 1.5 million H&E slides, excels in pan-cancer detection [Vorontsov et al., 2024]. UNI, pretrained on 100 million image patches from 100,000 whole slide images across 20 tissue types, outperforms previous models on various pathology tasks [Chen et al., 2024]. Prov-GigaPath, trained on 1.3 billion image tiles from 171,000 slides, demonstrates strong performance in biomarker prediction and pan-cancer detection [Xu et al., 2024]. All three models use variations of the DINO v2 algorithm for pretraining [Oquab et al., 2023].

2.2 EGFR Status Prediction

Several studies have demonstrated the feasibility of MIL to predict EGFR mutations from H&E stained WSIs. Coudray et al. were among the first to train a convolutional neural network (CNN) on TCGA data for EGFR prediction in NSCLC patients [Coudray et al., 2018]. Their approach involved training an CNN on patches from manually curated regions of interest, followed by aggregating patch-level predictions to make slide-level classifications. However, their reliance on manual curation and a dataset with high tumor purity limited clinical applicability. Sethi et al. developed an approach that used self-supervised contrastive learning to pretrain a feature extractor, followed by attention-based MIL to aggregate patch features for slide-level prediction without requiring manual annotations [Sethi et al., 2021]. Similarly, [Pao et al., 2023] used a CNN to extract patch features, which were then aggregated using an attention mechanism to produce slide-level predictions. While these studies have shown promising results, several limitations remain. Most models have been developed and validated primarily on surgical resection specimens, with limited evaluation on biopsy samples that are more common in clinical practice. Additionally, there are no studies, to the best of our knowledge, that have shown how to integrate these models in clinical practice for real-time EGFR status assessment.

3 Proposed Methods

This papers proposes the following two approaches to predict EGFR mutation status directly from H&E stained WSIs:

Approach 1: This approach consists of training two separate models –

Patch-level encoder: This encoder extracts a feature vector for each patch of a given WSI. The encoder's weights are frozen after training, and it serves as a feature extractor for the next stage.

Slide-level aggregator: This model takes as input the feature vectors produced by the patch-level encoder for all patches in a WSI. It then aggregates these features to produce a slide-level prediction of EGFR mutation status.

Figure 1a shows an illustration of the proposed Approach 1. During patch-level encoder training, tumor containing region of interest (ROI) is extracted for each WSI in the training set by using Otsu's thresholding [Otsu et al., 1975] followed by morphological dilation for background and small (< 64 pixels) object (tissue area) removal. Non-overlapping patches of size 224×224 pixels are then extracted from the identified ROI for every WSI. These patches are subsequently used as input to a vision transformer (ViT) [Dosovitskiy et al., 2021] that learns to assign each input patch to one of the histological types of lung cancer – acinar, solid, lepidic, papillary, or micropapillary [Nicholson et al., 2022]. The ground truth histological labels for each WSI are obtained by extracting the results of the comprehensive genomics testing and curating the EGFR mutations that are biologically relevant and targets of mutation specific therapy. The weights of the ViT are frozen after training and it is used as an encoder to extract feature vectors (CLS token [Dosovitskiy et al., 2021]) for input patches of size 224×224 pixels in the next stage.

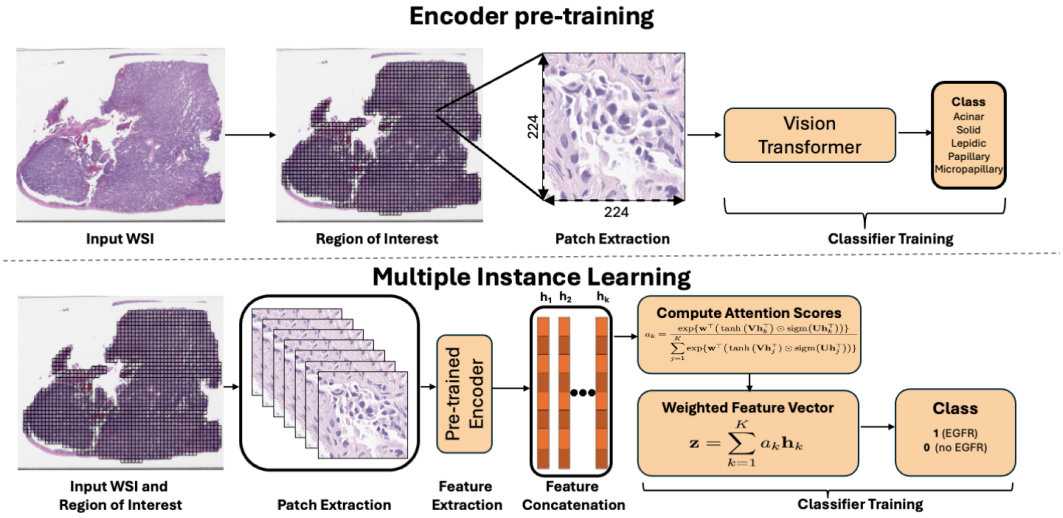
Slide-level aggregator obtains feature vectors for every (non-overlapping) patch obtained from an ROI in a given WSI and applies MIL pooling on these patch-level feature vectors to compute a slide-level feature vector, which is subsequently used for EGFR status prediction. In this paper, gated multi-head attention (GMA) as proposed in [Ilse et al., 2018] is used for slide-level aggregation. Note that the weights of slide-level aggregator are only updated in the second stage of Approach 1 while the pre-trained encoder weights are frozen as illustrated in 1a.

Approach 2: In this proposed approach the encoder weights are no longer frozen and are updated along with the MIL weights as illustrated in Figure 1b. Noticeably, this approach is computationally expensive as it requires processing multiple WSIs during training that involves ROI identification, patch extraction, a forward pass through the encoder followed by MIL mechanism, loss function computation, and then backpropagation to update the weights of both the encoder and the MIL module.

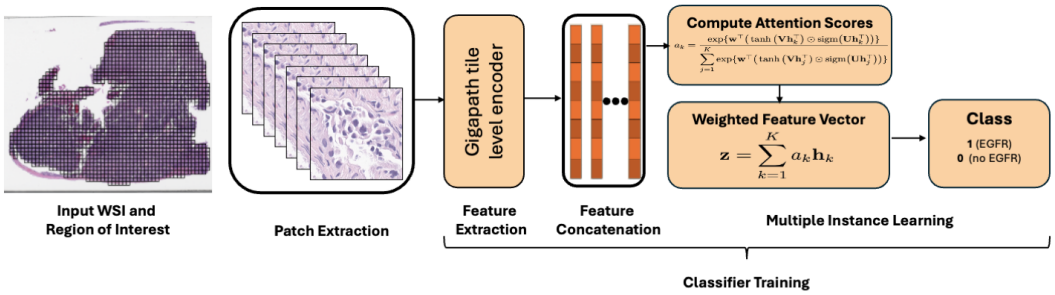
Thus, to implement this approach we adopted a distributed training method presented in [Campanella et al., 2024b] and summarized in Algorithm A1. It utilizes 2 or more GPUs during training[2], with one GPU dedicated to the aggregator process and N GPUs serving as encoder processes. For each slide, the algorithm distributes batches of image patches across the N encoder GPUs. These GPUs perform parallel forward passes through the encoder to generate patch-level features. The resulting features are then gathered on the aggregator GPU, where they are concatenated and processed by the aggregator to produce a slide-level feature vector. The slide level feature vector is used to predict the EGFR status, and a loss is computed. The backward pass begins on the aggregator GPU, computing gradients up to the concatenated features. These gradients are then split and scattered back to the encoder GPUs. Each encoder GPU computes a pseudo-loss using its portion of the gradients and backpropagates through the encoder. Finally, the encoder and aggregator models are updated using their respective gradients. This distributed approach allows for efficient processing of large WSIs by leveraging multiple GPUs, potentially improving training speed and enabling the handling of high-resolution image data. Note that the proposed Approach 2 does not require a surrogate task of lung histology classification to train an encoder.

In Real Time EGFR Assesment: Figure A1 illustrates the proposed in-real-time (IRT) pipeline to identify and process WSIs (using proposed Approach 1 or 2) for EGFR prediction in a live clinical setting. The purpose of the IRT pipeline is that, unlike publications that work with retrospective sample for which models can be built that overfit these datasets, the proposed methods must generalize such that any new WSI will perform as it would in a live clinical scenario. *Our institute* processes $90 - 110$ NSCLC cases per month for which EGFR testing is clinically indicated. The proposed IRT pipeline identifies slides that are scanned for molecular testing as well as the slides scanned from the same surgical pathology block for which molecular testing was ordered. Two watchers run every hour to identify 1) which slides have been scanned; 2) which lung cancer cases are sent for molecular analysis. When a slide that matches the molecular case is identified, the slide is transferred from the

[2]Inference is performed using a single GPU after training the model with Approach 2.



(a) Proposed Approach 1: A two-step process for predicting EGFR mutation status. Step 1: An encoder (vision transformer [Dosovitskiy et al., 2021] is shown as example) is trained on the surrogate task of lung histology classification. Step 2: The pre-trained encoder with frozen weights is used as a feature extractor for patches, followed by a slide-level aggregator (gated attention [Ilse et al., 2018] is shown as example) to predict EGFR mutation status.



(b) Proposed Approach 2: An end-to-end training process for predicting EGFR mutation status that updates both the encoder and MIL weights simultaneously.

Figure 1: The proposed approaches for predicting EGFR mutation status from H&E stained WSIs.

archive and the AI model is run immediately. This allows for a real-time EGFR prediction. If two or more WSIs are scanned, the mean probability score is used. When there is a large deviation in score from one slide to the next, the case is flagged to assess if one of the slides is from a different block than molecular testing.

4 Experiments and Results

4.1 Data Description

We applied the proposed approaches on a real-world clinical data to predict EGFR mutation status consisting of digitized slides of LUAD patient – scanned $20\times$ magnification with Aperio AT2 digital slide scanner from Leica Biosystems, at *our institute* paired with ground truth EGFR mutational status obtained from the IMPACT sequencing panel [Cheng et al., 2017]. Additionally, we evaluated the performance of the proposed methods on an external dataset from the TCGA LUAD cohort [Network et al., 2014]. Table A1 presents the distribution of WSIs across different stages of our study for two distinct approaches. Note that while Approach 2 utilizes all slides in the training set, the training set is divided into two subsets: one for encoder pre-training and another for MIL

model training for Approach 1. Both approaches share identical model selection (validation) and independent testing sets, ensuring a fair comparison. All slides in the training, model selection, and independent test sets were scanned at $20\times$ magnification. For the IRT set, both $20\times$ and $40\times$ lung cancer slides (requiring EGFR testing) scanned at *our institute* in the past 90 days are used. For $40\times$ slides in the IRT set, 448×448 sized patches are extracted and downsampled to 224×224 for a forward pass through the trained models of both approaches.

4.2 Results

The learning parameters for the two proposed approaches are presented in Table A2 in Appendix A. EGFR status prediction performance of all experiments is reported in terms of the area under curve (AUC), sensitivity, and specificity metrics for binary classification (EGFR positive or negative with 0.5 threshold). The best model for each approach is selected on the basis of the best AUC value on the model selection (validation) set (see Table A1).

Table A3 in Appendix A presents the results of EGFR status prediction with the proposed Approach 1 using different encoder architectures with GMA [Ilse et al., 2018] on the validation set. Table A3 shows that ViT-B [Dosovitskiy et al., 2021] achieved the highest AUC among all encoders. While ViT-L [Dosovitskiy et al., 2021] shows marginally better performance, the improvement does not justify its increased parameter count. Consequently, ViT-B is used as the encoder with GMA based MIL for all subsequent experiments with Approach 1. Prov-Gigapath [Xu et al., 2024] is used as the encoder model for fine-tuning with GMA based MIL in Approach 2 as it has shown better performance for several downstream tasks [Campanella et al., 2024a].

Results in Table 1 reveal that while Approach 1 marginally outperformed on the model selection set, Approach 2 demonstrated superior generalization, consistently achieving higher AUC values on both the independent test set and IRT cases across $20\times$ and $40\times$ magnifications. This enhanced performance of Approach 2 can be attributed to two key factors: (1) the use of the Prov-Gigapath foundation model, pre-trained on a large corpus of pathology images, providing more robust initial weights, and (2) the end-to-end learning paradigm that allows fine-tuning of encoder weights based on EGFR prediction errors in the training set. These advantages enable Approach 2 to capture intricate pathological features more effectively and adapt its feature extraction process specifically to the EGFR prediction task. Consequently, Approach 2 shows greater potential for EGFR mutation prediction in clinical settings, despite being more expensive to train than Approach 1.

Table 1: Performance metrics for the proposed approaches across different datasets at 0.5 threshold for binary classification. Bold text indicates the best model (in terms of AUC).

Dataset	**Approach 1**			**Approach 2**		
(n = # of WSIs)	**AUC**	**Sensitivity**	**Specificity**	**AUC**	**Sensitivity**	**Specificity**
Model Selection (n = 260)	**0.96**	0.83	0.92	0.90	0.88	0.93
Independent Testing (n = 6300)	0.89	0.72	0.96	**0.90**	0.84	0.95
TCGA-LUAD (n = 519)	0.78	0.60	0.81	**0.86**	0.78	0.74
IRT (n = 1000)	0.80	0.65	0.80	**0.83**	0.62	0.83

For IRT, the inference time for both approaches is about 2 minutes per slide, offering a significant improvement in turnaround time compared to conventional methods. The most common *Idylla*^{TM} rapid (PCR-based) EGFR testing has a 2-day turnaround time while suffering from issues with technical sensitivity, potentially missing important mutations. In contrast, the more comprehensive IMPACT-based genomic assessment [Cheng et al., 2017], has a 18-days turnaround time and is more costly. The proposed method offer a significantly shorter turnaround time, processing all slides scanned in a single day at the hospital within 30-45 minutes (median turn around time is 20 minutes).

5 Conclusion

This paper demonstrates the feasibility and clinical value of AI-driven EGFR mutation prediction from histopathology images in NSCLC patients. The proposed approaches offer rapid turnaround (2 minutes per slide), cost-effectiveness, and performance comparable to rapid PCR-based tests. The in-real-time (IRT) pipeline shows promise for expediting mutation detection, potentially leading to faster treatment decisions and optimized tissue sample utilization. As models are refined based on IRT performance, further improvements are anticipated. This work represents a significant step towards integrating AI-powered diagnostics into NSCLC management, advancing precision oncology.

References

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

Gabriele Campanella, Shengjia Chen, Ruchika Verma, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, Kuan-lin Huang, Ricky Kwan, et al. A clinical benchmark of public self-supervised pathology foundation models. *arXiv preprint arXiv:2407.06508*, 2024a.

Gabriele Campanella, Eugene Fluder, Jennifer Zeng, Chad Vanderbilt, and Thomas J Fuchs. Beyond multiple instance learning: Full resolution all-in-memory end-to-end pathology slide modeling. *arXiv preprint arXiv:2403.04865*, 2024b.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

Donavan T Cheng, Meera Prasad, Yvonne Chekaluk, Ryma Benayed, Justyna Sadowska, Ahmet Zehir, Aijazuddin Syed, Yan Elsa Wang, Joshua Somar, Yirong Li, et al. Comprehensive detection of germline variants by msk-impact, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC medical genomics*, 10:1–9, 2017.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Feny"o, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Jack Grant, Anne Stanley, Kevin Balbi, Gareth Gerrard, and Philip Bennett. Performance evaluation of the biocartis idylla EGFR mutation test using pre-extracted DNA from a cohort of highly characterised mutation positive samples. 75(4):241–249, 2022. ISSN 1472-4146. doi: 10.1136/jclinpath-2020-207338.

Crispin T Hiley, John Le Quesne, George Santis, Rowena Sharpe, David Gonzalez de Castro, Gary Middleton, and Charles Swanton. Challenges in molecular testing in non-small-cell lung cancer patients with advanced disease. *The Lancet*, 388(10048):1002–1011, 2016.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pages 2127–2136. PMLR, 2018.

Amir Momeni-Boroujeni, Paulo Salazar, Tao Zheng, Nana Mensah, Ivelise Rijo, Snjezana Dogan, JinYuan Yao, Christine Moung, Chad Vanderbilt, Jamal Benhamida, Jason Chang, William Travis, Natasha Rekhtman, Marc Ladanyi, Khedoudja Nafa, and Maria E. Arcila. Rapid EGFR mutation detection using the idylla platform: Single-institution experience of 1200 cases analyzed by an in-house developed pipeline and comparison with concurrent next-generation sequencing results. 23(3):310–322, 2021. ISSN 1943-7811. doi: 10.1016/j.jmoldx.2020.11.009.

Cancer Genome Atlas Research Network et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543, 2014.

National Comprehensive Cancer Network. Nccn clinical practice guidelines in oncology: Non-small cell lung cancer. https://www.nccn.org/professionals/physician_gls/pdf/nscl.pdf, 2024. Version 2.2024, Accessed: 2024-08-30.

Andrew G Nicholson, Ming S Tsao, Mary Beth Beasley, Alain C Borczuk, Elisabeth Brambilla, Wendy A Cooper, Sanja Dacic, Deepali Jain, Keith M Kerr, Sylvie Lantuejoul, et al. The 2021 who classification of lung tumors: impact of advances since 2015. *Journal of Thoracic Oncology*, 17(3):362–387, 2022.

Nicola Normanno, Kathi Apostolidis, Audrey Wolf, Raed Al Dieri, Zandra Deans, Jenni Fairley, Jörg Maas, Antonio Martinez, Holger Moch, Søren Nielsen, et al. Access and quality of biomarker testing for precision oncology in europe. *European journal of cancer*, 176:70–77, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11 (285-296):23–27, 1975.

James J Pao, Mikayla Biggs, Daniel Duncan, Douglas I Lin, Richard Davis, Richard SP Huang, Donna Ferguson, Tyler Janovitz, Matthew C Hiemenz, Nathanial R Eddy, et al. Predicting egfr mutational status from pathology images using deep learning and histopathologic features on whole-slide images. *Scientific Reports*, 13(1):4404, 2023.

Suresh S. Ramalingam, Johan Vansteenkiste, David Planchard, Byoung Chul Cho, Jhanelle E. Gray, Yuichiro Ohe, Caicun Zhou, Thanyanan Reungwetwattana, Ying Cheng, Busyamas Chewaskulyong, Riyaz Shah, Manuel Cobo, Ki Hyeong Lee, Parneet Cheema, Marcello Tiseo, Thomas John, Meng-Chih Lin, Fumio Imamura, Takayasu Kurata, Alexander Todd, Rachel Hodge, Matilde Saggese, Yuri Rukazenkov, and Jean-Charles Soria. Overall survival with osimertinib in untreated, *EGFR* -mutated advanced NSCLC. 382(1):41–50, 2020. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa1913662. URL http://www.nejm.org/doi/10.1056/NEJMoa1913662.

Gregory J Riely, Douglas E Wood, David S Ettinger, Dara L Aisner, Wallace Akerley, Jessica R Bauman, Ankit Bharat, Debora S Bruno, Joe Y Chang, Lucian R Chirieac, et al. Non–small cell lung cancer, version 4.2024, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 22(4):249–274, 2024.

Richard L Schilsky and Dan L Longo. Closing the gap in cancer genomic testing. *N. Engl. J. Med*, 387:2107–2110, 2022.

Amit Sethi, Lisa Sha, Ravi Das, and Deepak Kumar. Artificial intelligence-driven image analysis for cancer detection and prognosis. *Nature Reviews Cancer*, 21(9):546–558, 2021.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, pages 1–12, 2024.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024.

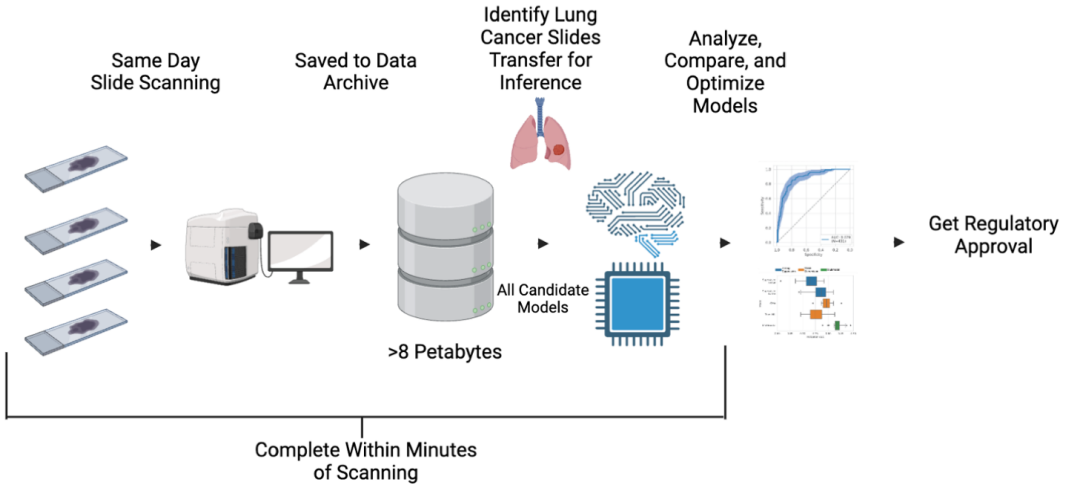


Figure A1: In-real-time (IRT) pipeline for EGFR prediction involves scanning and storing WSIs, retrieving slides of lung cancer patients that require EGFR testing, processing slides for EGFR prediction, and assessing results. The last step involves applying for regulatory approvals after completing required evaluations.

A Appendix

Table A1: Number of WSIs in the training, model selection (validation), and testing sets

	Encoder Pre-training	**MIL Model Training**	**Model Selection**	**Independent Testing**	**TCGA**	**IRT**
Approach 1	3475	1558	260	6300	514	1000
Approach 2	5033					

Table A2: Training parameters for the proposed approaches

Parameter	**Approach 1**		**Approach 2**
	Encoder Pre-training	**MIL Training**	
Number of GPUs	1	1	4
GPU	NVIDIA A100		
Number of epochs (N_{epochs})	80	50	150
Learning rate (η)	1×10^{-5}	1×10^{-4}	
Optimizer	Adam		
Weight decay (λ)	0.01	1×10^{-4}	1×0.001
Loss function	Cross-Entropy	Weighted Binary Cross-Entropy	Weighted Binary Cross-Entropy
Learning rate scheduler	Cosine with restarts	StepLR ($N_{\text{step}} = 15$, $\gamma = 0.1$)	Cosine with restarts
Class weighting	Not used	$w_{\text{positive}} = 0.7$	$w_{\text{positive}} = 0.7$

Table A3: Performance metrics of various encoders with MIL using GMA [Ilse et al., 2018] on the validation set (260 slides, Table A1)

Encoder Architecture	**AUC**	**Sensitivity**	**Specifity**
ResNet50	0.82	0.76	0.80
ViT-S	0.84	0.79	0.81
ViT-B	0.96	0.83	0.92
ViT-L	0.97	0.85	0.91

Algorithm A1 Distributed Encoder-Aggregator Training

Input: $N + 1$ GPUs, WSI Dataset D, Encoder E, Aggregator A
Output: Trained Encoder E and Aggregator A

Initialize:
  Rank 0 GPU as aggregator process
  Ranks 1 to N GPUs as encoder processes
  Wrap E with DistributedDataParallel

for each epoch **do**
  for each slide in D **do**
    Distribute N batches of image patches to N encoder GPUs using distributed sampler
    ▷ Forward Pass
    for rank r in 1 to N **in parallel do**
      $f_r \leftarrow E(\text{batch}_r)$ ▷ Generate patch-level features
    end for
    $F \leftarrow \text{Gather}(f_1, \ldots, f_N)$ to rank 0 ▷ Break computing graph

    On rank 0:
      $F_{\text{concatenated}} \leftarrow \text{Concatenate}(F)$
      slide_feature $\leftarrow A(F_{\text{concatenated}})$ ▷ Slide-level aggregator
      output $\leftarrow \text{Project}(\text{slide\_feature})$ ▷ EGFR status prediction
      loss $l \leftarrow \text{ComputeLoss}(\text{output}, \text{ground\_truth})$
      ▷ Backward Pass
    On rank 0:
      Compute gradients from l back to $F_{\text{concatenated}}$
      Split gradients into N chunks: $g_1, \ldots, g_N$
      Scatter$(g_1, \ldots, g_N)$ to ranks 1 to N

    for rank r in 1 to N **in parallel do**
      pseudo_loss $l_e \leftarrow N \times \sum_{i=1}^{|F|}(f_r[i] \times g_r[i])$
      Backpropagate l_e through E
    end for
    ▷ Update
    Update E on ranks 1 to N (handled by DistributedDataParallel)
    Update A on rank 0
  end for
end for

return E, A