GMILT: A Novel Transformer Network That Can Noninvasively Predict EGFR Mutation Status

Wei Zhao, Weidao Chen, Ge Li, Du Lei, Jiancheng Yang, *Member, IEEE*, Yanjing Chen, Yingjia Jiang, Jiangfen Wu, Bingbing Ni, Yeqi Sun, Shaokang Wang, Yingli Sun, Ming Li, and Jun Liu

***Abstract*— Noninvasively and accurately predicting the epidermal growth factor receptor (EGFR) mutation status is a clinically vital problem. Moreover, further identifying the most suspicious area related to the EGFR mutation status can guide the biopsy to avoid false negatives. Deep learning methods based on computed tomography (CT) images may improve the noninvasive prediction of EGFR mutation status and potentially help clinicians guide biopsies by visual methods. Inspired by the potential inherent links between EGFR mutation status and invasiveness information, we hypothesized that the predictive performance of a deep learning network can be improved through extra utilization of the invasiveness information. Here, we created a novel explainable transformer network for EGFR classification named gated multiple instance learning transformer (GMILT) by integrating multi-instance learning and discriminative weakly supervised feature learning. Pathological invasiveness information was first introduced into the multitask model as embeddings. GMILT was trained and validated on a total of 512 patients with adenocarcinoma and tested on three datasets (the internal test dataset, the external test dataset, and The Cancer Imaging Archive (TCIA) public dataset). The performance (area under the curve (AUC) = 0.772 on the internal test dataset) of GMILT exceeded that of previously published methods and radiomics-based methods (i.e., random forest and support vector machine) and attained a preferable generalization ability (AUC = 0.856 in the TCIA test dataset and AUC = 0.756 in the external dataset). A diameter-based subgroup analysis further verified the efficiency of our model (most of the AUCs exceeded 0.772) to noninvasively predict EGFR mutation status from computed tomography (CT) images. In addition, because our method also identified the "core area" of the most suspicious area related to the EGFR mutation status, it has the potential ability to guide biopsies.**

***Index Terms*— Epidermal growth factor receptor (EGFR), computed tomography, multiple instance learning (MIL), transformer.**

Manuscript received 25 March 2022; revised 15 June 2022; accepted 8 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 82102157 and Grant 61976238, in part by the Hunan Provincial Natural Science Foundation for Excellent Young Scholars under Grant 2022JJ20089, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2021JJ40895, in part by the Research Project of Postgraduate Education and Teaching Reform of Central South University under Grant 2021JGB147 and Grant 2022JGB117, in part by the Clinical Research Center For Medical Imaging in Hunan Province under Grant 2020SK4001, and in part by the Science and Technology Innovation Program of Hunan Province under Grant 2021RC4016. *(Wei Zhao and Weidao Chen are co-first authors.) (Corresponding authors: Ming Li; Jun Liu.)*

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of The Second Xiangya Hospital under Approval No. 2022k012, and performed in line with the Council for International Organizations of Medical Sciences (CIOMS) guideline.

Wei Zhao is with the Department of Radiology, The Second Xiangya Hospital, Central South University, Changsha 410011, China, and also with the Clinical Research Center for Medical Imaging in Hunan Province, Central South University, Changsha 410011, China (e-mail: wei.zhao@csu.edu.cn).

Weidao Chen, Jiangfen Wu, and Shaokang Wang are with the International Center, InferVision, Beijing 100020, China (e-mail: chenweidao163@163.com; wjfyunzhu@163.com; wshaokang@infervision.com).

Ge Li is with the Department of Radiology, Xiangya Hospital, Central South University, Changsha 410008, China (e-mail: ligeanyi@126.com).

Du Lei is with the Department of Psychiatry and Behavioral Neuroscience, College of Medicine, University of Cincinnati, Cincinnati, OH 45221 USA (e-mail: du.lei@kcl.ac.uk).

Jiancheng Yang and Bingbing Ni are with the Department of Electronic Engineering and the SJTU-UCLA Joint Center for Machine Perception and Inference, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jekyll4168@sjtu.edu.cn; nibingbing@sjtu.edu.cn).

Yanjing Chen and Yingjia Jiang are with the Department of Radiology, The Second Xiangya Hospital, Central South University, Changsha 410011, China (e-mail: 962409428@qq.com; 544979495@qq.com).

Yeqi Sun is with the Department of Pathology, The Second Xiangya Hospital, Central South University, Changsha 410011, China (e-mail: zoesyq19831@csu.edu.cn).

Yingli Sun and Ming Li are with the Department of Radiology, Huadong Hospital Affiliated to Fudan University, Shanghai 200040, China (e-mail: sunyingli208ok@126.com; minli77@163.com).

Jun Liu is with the Department of Radiology, The Second Xiangya Hospital, Central South University, Changsha 410011, China, also with the Department of Radiology Quality Control Center, Central South University, Changsha 410011, China, and also with the Clinical Research Center for Medical Imaging in Hunan Province, Central South University, Changsha 410011, China (e-mail: junliu123@csu.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TNNLS.2022.3190671.

Digital Object Identifier 10.1109/TNNLS.2022.3190671

I. Introduction

NONSMALL cell lung cancer (NSCLC) accounts for more than 80% of lung cancer cases, and lung adenocarcinoma is the most common type of NSCLC, with a five-year relative survival rate of 5% for patients diagnosed with metastatic disease [1]. The advancement of genomics and precision medicine has facilitated the development of cancer treatment paradigms, such as targeted therapy with epidermal growth factor receptor (EGFR)-tyrosine kinase inhibitors (TKIs) [2].

One study found that EGFR-mutant patients have an objective response rate (ORR) of approximately 80% when treated with EGFR-TKIs [3]. In contrast, the administration of the EGFR-TKI gefitinib has no effect or results in even worse progression-free survival (PFS) and unnecessary costs when applied to patients without EGFR mutations [4]. Moreover, patients with EGFR mutations who therefore lack an inflammatory microenvironment have an unfavorable ORR to immune checkpoint inhibitor (ICI) treatments [5]. Therefore, an accurate pretreatment estimation of EGFR mutation status could significantly help clinicians select eligible patients for EGFR-TKI treatment, thus supporting individualized decision-making and improving patient outcomes to the greatest extent possible.

Informative tissue-based assays, such as next-generation sequencing (NGS), remain the standard medical procedure to determine EGFR mutation status. However, the inherent disadvantages of these approaches, such as the need for invasive procedures, sampling bias, and high cost, limit their clinical application in some scenarios. Furthermore, EGFR mutation status and the immune landscape may change during cancer progression and/or therapy [6]. In clinical, different tumor sizes face different clinical issues in EGFR mutation testing. For example, tumors with sizes less than 1 cm might not have residual tissues for EGFR testing after histopathologic analysis. Another case is that patients with unresectable tumors may require repeated sampling to identify the EGFR mutation status and guide EGFR-TKI therapy. Therefore, highly efficient, noninvasive, longitudinal, high-throughput methods for predicting EGFR mutation status, preferably size-based subgroup analysis, are urgently needed in the clinic.

Recently, an increasing number of investigators have suggested that CT images contain rich information that may intrinsically reflect the inherent characteristics of EGFR mutation status. Therefore, many investigators have attempted to achieve noninvasive EGFR identification using imaging data mining methods (i.e., radiomics and deep learning) [7]–[9]. However, radiomics involves time-consuming image segmentation and inevitable feature selection, making it difficult to apply in the clinical environment. In contrast, deep learning can substantially overcome the abovementioned disadvantages and outperform radiomics methods in the same task [8], [10], [11]. Nonetheless, mining CT images to efficiently predict EGFR mutation status by deep learning remains a substantial challenge, notably because previous studies only conducted a single task and ignored the inherent correlations between mutation status and other biological behaviors that may influence the mutation status of a gene, such as pathological categories and the internal microenvironment.

Recently, several scientific papers reported that EGFR mutations can occur in the early stage of lung adenocarcinoma and during tumor initiation from preneoplastic to neoplastic lung parenchyma conditions, including atypical adenomatous hyperplasia (AAH), adenocarcinoma *in situ* (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC) [6], [12]. In addition, the probability of mutation can potentially increase as the invasive extent progresses [13]. Thus, it is reasonable to hypothesize that invasiveness information can have predictive value in evaluating EGFR mutation status. To the best of our knowledge, no previous studies have investigated the hypotheses to date. Multitask learning aims to improve such generalization by leveraging domain-specific information contained in the training signals of related tasks and has shown promising results w.r.t. performance, computations, and/or memory footprint, by jointly tackling multiple tasks through a learned shared representation [14]. Recently, a new kind of encoder–decoder neural architecture, transformer, is proposed, which can effectively extract and utilize the relational features between different input data or feature representations [15], [16]. Incorporating multitask learning and transformer may improve the predictive performance by mining the relational patterns between invasiveness and EGFR mutation status. Notably, not all lesion sections reveal EGFR expression simultaneously due to the inherent heterogeneity of the tumor, which results in the lack of valuable features extracted from the lesion patches and can easily lead to overfitting. Therefore, improving the utilization of valuable features and increasing the signal-to-noise ratio (SNR) of feature space should be addressed. Multiple-instance learning methods integrating the attention mechanism can improve the feature expression ability of the model while reducing the data annotation requirements [17]–[20]. In addition, the active learning method, which seeks to find the most informative samples in the model development process, can improve the model training efficiency and the model generalization ability [21].

Therefore, to improve the predictive performance of EGFR mutation status and investigate whether adding the invasiveness information of lung adenocarcinoma to the model could obtain better performance, we propose a novel gated multiple instance learning transformer (GMILT) architecture by integrating multiple-instance learning and active learning (see Fig. 1), which could efficiently exploit and utilize rich discriminative patterns by bridging the representation gap in the spatial and global information domains in the tumor. The main contributions are given as follows. First, we proposed an innovative online sample selection method to improve the generalizability of this model in an active learning setting. To further enhance the feature representation ability of visual transformers, we first applied the group ensemble method to incorporate cardinality constraints on visual words in each minibatch to leverage possible prior knowledge based on multitask learning, i.e., EGFR classification and pathological invasiveness classification. To the best of our knowledge, this is the first study to investigate the interaction effects between EGFR mutation status and invasiveness information with a deep learning model, and it is also the first study to introduce the transformer method to medical tasks to improve the efficiency of feature learning. Second, we validated our model on three different datasets and compared it to previous related deep learning studies to verify the advantages of the proposed framework. Finally, we designed an attention pooling that can be used to visualize model decisions and to provide precision guidance for biopsy procedures.

We collected 726 lung adenocarcinoma patients with EGFR mutation testing from three datasets. All nodules were manually segmented, and labeled as EGFR mutant (EGFR+)/wild-type (EGFR-). An originally proposed GMILT model, incorporating multiple instance learning (MIL), transformer, active learning, and multitask learning algorithms, was constructed to efficiently utilize invasiveness information as embeddings to improve the performance of EGFR mutation status prediction and potentially guide biopsy.

II. Methods

This retrospective study was approved by The Second Xiangya Hospital, Institutional Review Board (IRB), which waived the requirement for informed consent.

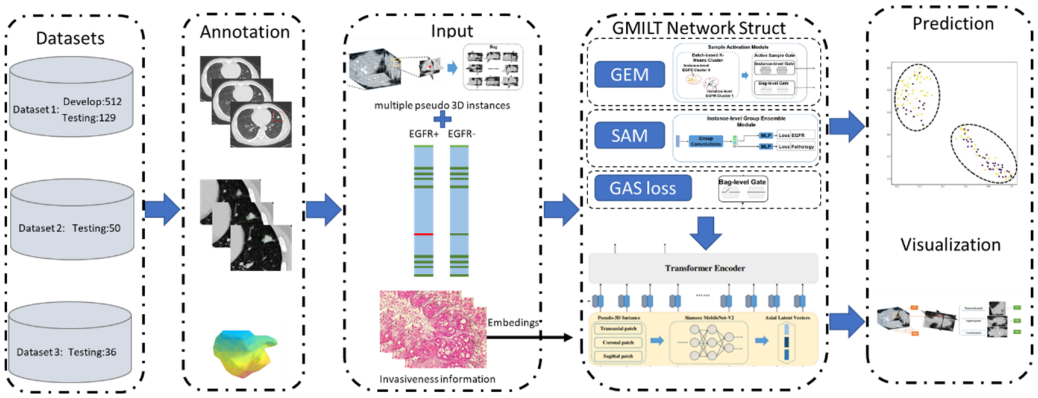


Fig. 1. Workflow of our study. We collected 726 lung adenocarcinoma patients with EGFR mutation testing from three datasets. All nodules were manually segmented and labeled as EGFR mutant (EGFR+)/wild-type (EGFR-). An originally proposed GMILT model, incorporating MIL, transformer, active learning, and multitask learning algorithms, was constructed to efficiently utilize invasiveness information as embeddings to improve the performance of EGFR mutation status prediction and potentially guide biopsy.

A. Patients and Inclusion Criteria

In this study, we collected three datasets for analysis. Dataset 1, including 640 patients from January 2013 to December 2018, was collected from Huadong Hospital and used for model development and validation. Dataset 2, including 50 patients from January 2020 to March 2021, was collected from Second Xiangya Hospital and used for external testing. Dataset 3, including 36 patients, was collected from The Cancer Imaging Archive (TCIA) public database [22] and used to validate the stability and generalization of the GMILT network. For datasets 1 and 2, the inclusion criteria: 1) patients who underwent thin-slice chest CT (0.75–1.5 mm) scans prior to biopsies or surgical treatment; 2) patients with detailed pathological reports diagnosing lung adenocarcinoma; and 3) patients with detailed EGFR mutation testing reports. The inclusion criteria for dataset 3 were given as follows: 1) CT images with slice thickness $\leq$1.5 mm (to avoid data inconsistency); 2) patients with EGFR mutations testing reports; 3) patients with pathology reports for the diagnosis of lung adenocarcinoma; and 4) those lesions that could be certainly identified as the resected or biopsied lesions. The exclusion criteria were: 1) CT images with slice thickness $>$1.5 mm; 2) patients without EGFR mutation testing reports; and 3) patients with pathological reports for the diagnosis other than lung adenocarcinoma. Only malignant nodules with EGFR testing results were included. The CT scanning and data preprocessing information were presented in Section 1 in the Supplementary Material.

B. Design of the Experiments and Model Construction

Considering the inherent relationship between the EGFR mutation status and pathological invasiveness information, we designed five deep learning models to comprehensively investigate the inherent interactions in a multitask environment, especially focusing on the incremental value of invasiveness information for predicting EGFR mutation status from different aspects (see Fig. 2).

C. Single Task Analysis

In this part, we constructed two single tasks to investigate the performance of deep learning in two separate tasks.

Model 1: Constructing a 2.5-D network to predict the invasiveness of lung adenocarcinoma without using EGFR mutational information.

Model 2: Proposing a novel MIL network (the baseline network, also called the baseline of GMILT) to predict the EGFR mutation status of lung adenocarcinoma without using pathological invasiveness information.

D. Interaction Analysis (Multitask)

In this part, we add the invasiveness information to model 2 in three different scenarios.

Model 3: Consider the invasiveness information as an **input** into the network to investigate its influence in predicting the EGFR mutation status.

Model 4: Consider the invasiveness information as **supervised information** to the network to simultaneously predict the EGFR mutation status and the invasiveness.

Model 5: Considering the invasiveness information as **semantical information embedded** into the proposed network to predict the EGFR mutation status of lung adenocarcinoma. In model 5, we first introduce two skills (embedding and active learning) to facilitate embedded intermediate feature learning, thus constructing the final proposed network—GMILT.

As the main purpose of the study was to noninvasively predict the EGFR mutation status, the performance of identifying the invasiveness of lung adenocarcinoma was not evaluated in our model 5.

E. Proposed Approach

In this section, we formulate the problem of EGFR type classification and describe our proposed GMILT approach. As illustrated in Fig. 3, compared to traditional MIL, GMILT first transforms a raw unseparated bag into multiple pseudo-3-D instances with aggregated semantic representation fusing the transaxial, coronal, and sagittal representations based on the transformer architecture. It then combines the deep pseudo-3-D instances into the bag representation using attention-based MIL pooling in the transformer encoder module. It finally transforms the bag representation into the final prediction by using a neural network to learn the Bernoulli

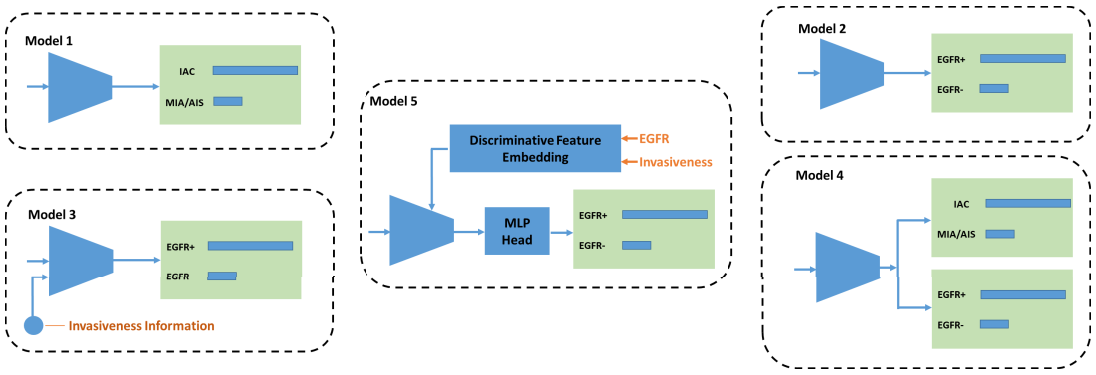


Fig. 2. Overall design of the five models. The MLP head contains two fully connected layers that are applied for feature transformation and nonlinearity.

distribution of the bag. Meanwhile, in the transformer encoder module, we propose the group ensemble module (GEM) and the sample activation module (SAM) in conjunction with active smoothing (GAS) loss to improve the discriminative feature learning and generalization ability of GMILT.

1) Problem Setting: In this part, we will first analyze the main challenges in EGFR type identification caused by tumor heterogeneity and multitask learning, and then provide corresponding countermeasures.

For the annotation of CT images, lesionwise labels directly come from the gene detection of EGFR, which is concretely based on the part of sampling through biopsy or surgery. In this sense, the diagnostic results of EGFR type are restricted by the sampled parts, and the results of the residual parts of lesions remain unknown (actually without the ground truth) even when this image has already been labeled. This poses a great challenge for the utilization of traditional supervised learning models. To address this challenge, we borrow an idea from an MIL framework, which is a typical weakly supervised learning paradigm to address patient-level (bag-level) prediction without knowledge of any region-level annotation.

In the MIL setting, one lesionwise CT cube is divided into N 3-D subparts of equal size, which are seen as instances. We consider a group of N instances (called "Bag") and assume that each group from a positive class sample contains at least a few 3-D instances with positive class-specific information, whereas each group from a negative sample does not contain any 3-D instances having positive class-specific information. However, in conflict with MIL constraints, bag-level false-negative results in gene detection still exist. To address this challenge, we introduce the active learning method [22] to control whether a sample is used to train the model online, which improves the training efficiency. Nevertheless, it is also worth noting that 3-D inputs with a low SNR are unfavorable for feature learning, and a combination of MIL and 3-D inputs leads to an increase in computation and memory costs, limiting the usage of minibatch sizes and the convergence ability of the backpropagation (BP) neural network model. To address this challenge, we implement an MIL-based visual transformer using pseudo-3-D inputs, which can preserve 3-D semantic representations and reduce the effects of noise and memory cost simultaneously. Another challenge is the interactive effects among multiple supervised tasks in the multitask learning setting. We noticed that most area under the curves (AUC) of previous CT image-based EGFR classification models varies in the range of 65% $\sim$ 81% [8], [10], showing the difficulty of feature learning. Thus, it is essential to construct a multitask learning approach that makes representations of EGFR mutation status more discriminative and promotes EGFR classification. To address this issue, we adopt the ensemble learning technique during training by structured coding in the process of feature learning with EGFR and pathology information. In other words, to address the aforementioned challenges, our proposed approach has four major components: 1) baseline GMILT; 2) SAM; 3) GEM; and 4) GAS loss. In particular, we build a deep MIL-based visual transformer with pseudo-3-D inputs to predict the bag level (i.e., EGFR+ or EGFR-). In the training stage, we incorporate the three semantic embedding modules—SAM, GEM, and GAS losses—into the baseline of the GMILT model.

2) Preprocessing: Deep Instance Generation: Considering a set of patients, $\{X_i\}$, $i = 1, \ldots, N$, each patient has an annotated lesionwise cube based on CT images with its EGFR mutational label $Gi \in \{0,1\}$ and pathological invasiveness label $G'i \in \{0,1\}$. We first crop the lesionwise cubes from the original CT scans according to the 3-D annotations (bounding boxes). Then, each lesionwise cube is divided into m equal parts, and based on their centroid points, we obtain the transaxial, coronal, and sagittal patches. Specifically, the transaxial, coronal, and sagittal patches, which pass through the same centroid point, make up an MIL instance representing the 3-D information of the responding equal part (called a pseudo-3-D instance). Finally, the m MIL instances constitute a patient-level MIL bag with a bag-level label (see Fig. 3).

3) Baseline of Gated MIL Transformer: The graphical representation of the proposed baseline of the GMILT model is illustrated in Figs. 4 and 5. GMILT receives as input a sequence of pseudo-3-D instances. In GMILT, we view a bag of pseudo-3-D instances as a visual sentence that represents the 3-D lesion, i.e., a visual sentence is composed of a sequence of visual words

$$X^i -> \left[x^{i,1}, x^{i,2}, \ldots, x^{i,m}\right] \tag{1}$$

where $x^{i,j}$ is the jth word visual word of the ith visual sentence, $j = 1, 2, \ldots, m$.

To handle pseudo-3-D instances, we prepend a learnable embedding on the sequence of transaxial, coronal, and sagittal patches with a Siamese MobileNet-V2 network outputting the

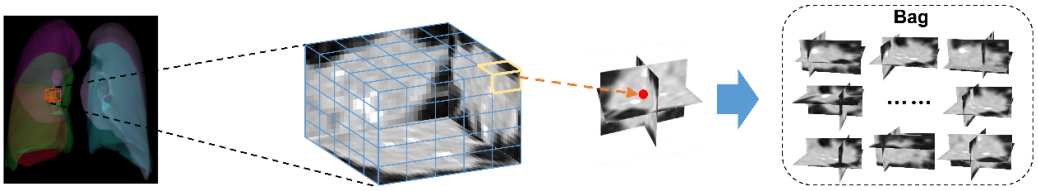


Fig. 3. Overview of deep instance generation. The origin CT volumes were first preprocessed using cropping to extract the lesionwise cubes. Then, the cubes were partitioned equally to construct the MIL instances.

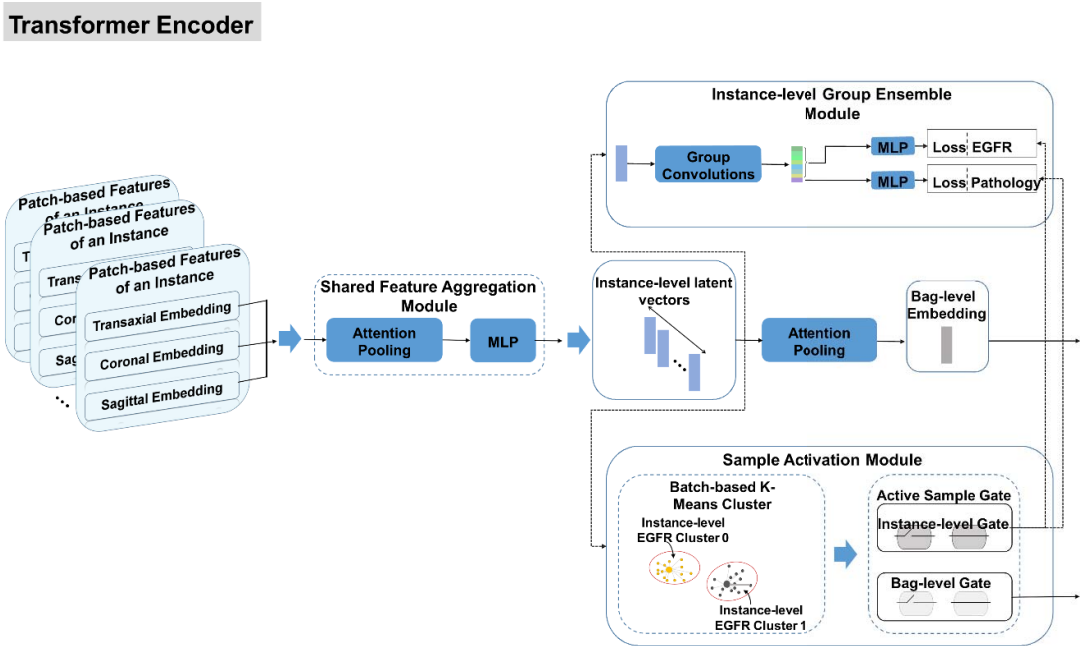


Fig. 4. Framework of the transformer encoder module. In this module, we first fused the patch-based features within an instance using a shared feature aggregation module, which is based on an attention mechanism, to obtain the instance-level latent vectors. Then, we used the attention pooling block to fuse the instance-level latent vectors, which belongs to the same bag, obtaining the bag-level embedding feature vectors. In addition, with the help of the K-means algorithm, the SAM clusters the instance-level latent vectors and obtains the gating values of the instance level and the bag level by comparing the clustering results with the real labels. Meanwhile, using instance-level latent vectors and instance-level gating values as input, the GEM uses group convolutions to perform structural encoding on the instance-level latent vectors.

corresponding patch embeddings. In this way, we transform the visual words into a sequence of word embeddings

$$Y^i = \left[y^{i,1}, y^{i,2}, \ldots, y^{i,m}\right] \tag{2}$$

where $\mathrm{y}^{i,j}$ is the jth word embedding, which contains a group of patch embeddings—($\mathrm{v}^{0,i,j}$, $\mathrm{v}^{1,i,j}$, $\mathrm{v}^{2,i,j}$), $\mathrm{v}^{0,i,j}$ is the transaxial patch embedding, $\mathrm{v}^{1,i,j}$ is the coronal patch embedding, and $\mathrm{v}^{2,i,j}$ is the sagittal patch embedding.

Clinical information embeddings (such as pathological invasiveness information) or position embeddings can be added to patch embeddings. The resulting sequence of latent vectors serves as input to the transformer encoder. In the baseline of GMILT, the transformer encoder consists of alternating layers of self-attention and FC layers, as shown in Figs. 4 and 5. We first use attention pooling and multilayer perception (MLP) to aggregate the patch features for each word, obtaining a single word representation $\Phi^{i,j}$ in the shared word embedding aggregation module. Subsequently, we implement a second attention pooling layer that performs attention-based permutation invariant pooling to obtain a single sentence representation z^i. Most notably, the two attention-based operations allow GMILT to learn both local information and 3-D global information in a visual word and across the visual sentence. Next, z^i is passed to the MLP head module to obtain predictions for the entire bag.

Attention Pooling: The self-attention operator is an interpretable symmetric function [23]. Formally, we denote $H = \{h1, \ldots, hN\}$ as the inputs with N embeddings. Then, the operator is defined as

$$z = \sum_{k=1}^{N} \alpha_k h_k \tag{3}$$

$$\alpha_k = \frac{\exp\left\{w^T \tanh\left(V h_k^T\right)\right\}}{\sum_{j=1}^{N} \exp\left\{w^T \tanh\left(V h_j^T\right)\right\}} \tag{4}$$

where $w \in R^{N\times 1}$ and $V \in R^{N\times D}$ are trainable parameters. In addition, α_k is considered the attention score per input embedding, indicating its contribution to the drawn conclusion, which is helpful for interpreting the trained model, i.e., interpretable analysis for identifying the potential core area. Importantly, the processing flow of the transformer encoder implements attention in attention architecture on patch embeddings and word embeddings using the basic operator differentiating to the existing attention mechanism.

4) SAM: In the training phase, given the bags input by each minibatch, our goal is to assess the beneficial effect of word embeddings (instance-level latent vectors) for the

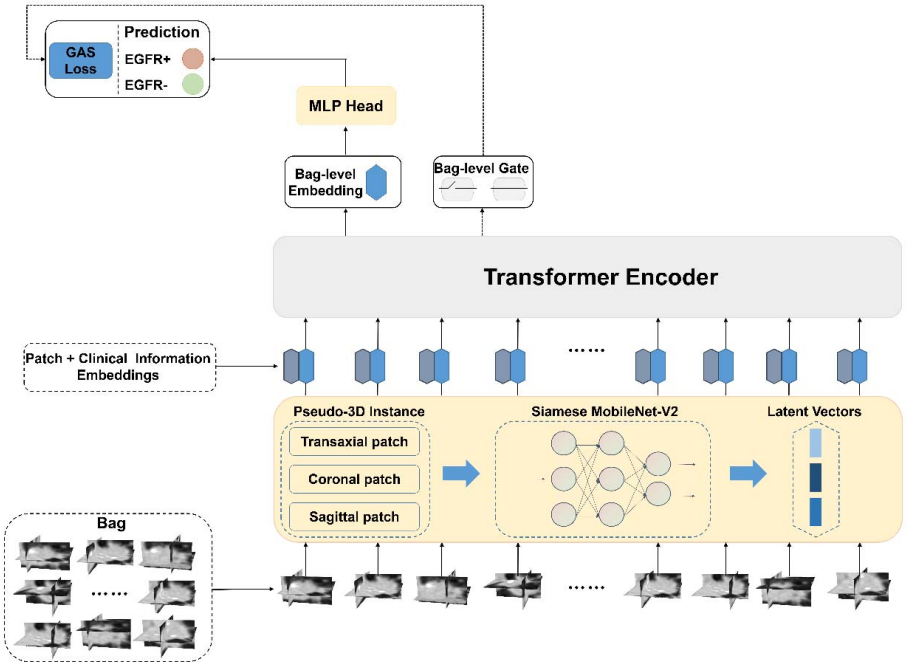


Fig. 5. Framework of the GMILT. First, based on the MIL setting, we used the generated "bag" of pseudo-3-D instances as input into the transformer-based model. Then, we used the Siamese MobileNet-V2 network to encode the multisectional patches from the three different 2-D view directions (i.e., transaxial, coronal, and sagittal axes) for each pseudo-3-D instance, obtaining the three corresponding latent vectors. Second, these latent vectors were fed to the transformer encoder module to obtain the bag-level embedding features. Third, the MLP head further integrated the bag-level feature vectors to output the predictive values of EGFR+/-. During training, GAS loss performed discriminative feature learning for EGFR classification using bag-level features and gating values.

discriminative learning of EGFR mutation status and define a sample weight to coordinate model training online with an adaptive gating technique. To achieve this, we infer that word embeddings with the same EGFR label should be matched to a single centroid that represents the EGFR type. Specifically, we set each batch containing the same number of positive and negative bags and perform K-means clustering on all instances in the batch to obtain two clusters using visual word embeddings x. Then, each instance's EGFR label is assigned by its bag's EGFR label. Based on the majority voting method, we mark the two clusters as the "EGFR−" and "EGFR+" clusters. Generally, in the MIL setting, the instances in a negative bag are largely negative. Therefore, the cluster, which contains most of negative instances, is firstly seen as "EGFR−" cluster; then the other cluster is seen as "EGFR+" naturally. On this basis, for each instance, if its cluster tag is the same as its EGFR label, its gate value is recorded as 1; otherwise, it is recorded as 0. Furthermore, for each bag, if the cluster labels of all instances in the bag are consistent with the EGFR label of the bag, its gate value is set to 1; otherwise, it is set to 0. Significantly, for positive (EGFR+) bags, its bag-level gate values are set to 1 because there are little fake positive bags based on the clinical scenario. Taking the gate values as the sample weights of the weighted loss function, we control whether the model performs supervised learning on the input samples online, as shown in Fig. 4. The formula is given as follows:

$$\delta(x) = \begin{cases} 1, & \text{label} == \text{label}_{\text{Kmeans}} \\ 0, & \text{otherelse} \end{cases} \tag{5}$$

where label denotes the input sample's real label based on annotation and $\text{label}_{\text{Kmeans}}$ denotes the input sample's clustering label by k-means.

5) GEM: As mentioned before, existing popular approaches of multitask learning with neural networks learn multiple related tasks simultaneously based on the underlying shared representation to improve their generalization ability. However, this approach neglects the impact of the complex relation between tasks in modeling, which may make multitask learning less efficient. GENet [24] is an efficient way to improve model capacity across a wide range of learning tasks. Here, we propose an instance-level semantic structured coding approach (i.e., GEM) to provide an extra regularization for the shared representation, as shown in Fig. 4. First, we apply group convolutions to each word embedding and divide the output vector into several groups. Then, we use a multihead structure to implement multitask learning by calculating the responding loss functions of the tasks simultaneously, i.e., EGFR classification and pathology classification. In particular, each group is one constituent member of our ensemble and has its own independent head classifier, which introduces diversity among the learning tasks. We conduct two tasks (EGFR and pathology) to train the model end to end simultaneously with its own objective function

$$\text{Loss}_{\text{GEM}} = \sum_{m=1}^{n} \text{Loss}_{e,m} + \text{Loss}_{p,m} \tag{6}$$

where Loss_{GEM} is the total objective function in this module, m is the group index, and n is the number of groups, i.e., $\text{Loss}_{m=1}$ and $\text{Loss}_{m=2}$ denote the instance-level loss functions of

EGFR classification and pathology classification, respectively. In particular, the loss function Loss_m is implemented with GAS loss, which is defined in the following section.

6) GAS Loss: To reduce the effect of noise labels while enhancing the discrimination of features, we design a loss function called GAS loss. The loss function draws on the idea of active learning and uses the gating values given by the SAM during forwarding propagation for online sample weighting to actively learn valuable samples. Meanwhile, the center constraint is applied to the feature space, which pushes the encoded features to the center of the cluster and reduces the differences within the class [25]. Finally, the weighted cross-entropy loss function and label smoothing algorithm are incorporated to further alleviate the impact of noise labels and balance the positive and negative samples. The loss function is given as follows:

$$\begin{aligned} L_{GAS} = \frac{1}{N}\sum_i -\big[\delta \cdot w_{\text{pos}}\hat{y}_i \log(p_i) + \delta \cdot w_{\text{neg}}\,(1-\hat{y}_i) \\ \times \log(1-p_i)\big] + \frac{1}{2}\left\|z_i' - c_{y_i}\right\|_2^2 \\ \hat{y}_i = \varepsilon(1-y_i) + (1-\varepsilon)y_i \end{aligned} \tag{7}$$

where N represents the number of training samples; yi represents the label of sample i, the positive class is set to 1, and the negative class is set to 0; pi is the probability that sample i is predicted to be a positive class; w_{pos} and w_{neg} are the weights of positive and negative samples, respectively, the value range is [0,1], and the values in this part are set to 0.75 and 0.25, respectively; ε is a small modulation parameter, and the value is set to 0.1 in this part; and ci denotes the yith class center of latent feature manifold $z'i$ and is updated based on the following feature distances [25].

Finally, with the previously defined loss functions, the overall objective to optimize our model (GMILT) can be formulated as

$$L_{\text{total}} = L_{\text{GEM}} + L_{\text{EGFR}} \tag{8}$$

where L_{EGFR} denotes the bag-level loss functions with the final prediction of GMILT based on L_{GAS}, as shown in Fig. 4.

7) Implementation Details: A total of 640 patients in dataset 1 were randomly divided into a development set ($n = 513$) and a test set ($n = 129$) at a 4:1 ratio. For training the proposed deep learning model, we use the Ranger [26] optimization with a batch size of 32 and a learning rate of 0.001. Ranger is a synergistic optimizer combining RAdam [27], LookAhead [28], and gradient centralization [29] in one optimizer, which accelerates convergence and reduces training difficulty. In addition, the Siamese MobileNet-V2 network is initialized using the pretrained parameter of ImageNet [30], and the dropout technique [31] is applied during training. To test the deep learning model, we select middle patches in the transaxial, coronal, and sagittal axes for each pseudo-3-D instance in the MIL setting to construct the input sample.

8) Visualization and Interpretation Analysis: For a given 3-D lesion, the two attention pooling layers assign importance values to each visual word, and the patches highlight the important spine regions. A new gradient-weighted class activation mapping (Grad-CAM) network was used to produce a coarse localization map highlighting the important regions in the image for predicting the concept [32].

9) Ablation Study: To assess the effect of bag size and group size on the GEM model during training on the proposed method, we performed two ablation studies to select the best hyperparameters k ($k = 5$) in MIL and μ ($\mu = 0.125$) in GEM (see Sections 2 and 3 in the Supplementary Material). Our deep learning model is implemented using the popular open-source framework PyTorch (1.6.0) and runs on an Nvidia GTX 1080Ti GPU.

F. Validation Analysis and Comparison Analysis With Related Publishers

We tested our proposed model on three datasets, including one internal dataset one external dataset, and a public dataset. To verify the efficiency of our newly proposed network, we also trained and tested three publishers' models [7], [8], [10] on our Dataset 1, and evaluated and compared the performance outcomes.

G. Validation Analysis, Comparison Analysis With Related Publishers, and Subgroup Analysis

In clinical scenarios, small lesions may have inadequate samples for gene analysis, and larger lesions without the indication for surgery may require repeated biopsies during systemic therapy. Given the different clinical demands and T-stage classification criterion, we further investigated the performance of our proposed model in four different subgroups categorized by the maximum diameter (MD) of the nodule ($0 < \text{MD} < 1$ cm; $1 \leq \text{MD} < 2$ cm; $2 \leq \text{MD} < 3$ cm; and 3 cm $\leq$ MD). Several subgroups had zero samples (i.e., the $0 < \text{MD} < 1$ cm subgroup in dataset 3 and the $1 \leq \text{MD} < 2$ cm subgroup in dataset 2), so we merged them with the adjacent subgroup and calculated the AUC.

H. Data Availability

The model code is available at https://github.com/TXVision/GMILT or on request to the corresponding author.

III. Results

A. Patients and Datasets

In Dataset 1, 640 patients were finally included and randomly divided into the model training dataset (383 patients), the validation dataset (129 patients), and the internal testing dataset (129 patients). Fifty patients from Dataset 2 were finally included and used for model external testing. Dataset 3 from the TCIA, finally including 36 patients, was used to validate the stability and generalization of the GMILT network as a public testing dataset. The distributions of patients were presented in Table I.

B. Model Construction

We constructed five models to comprehensively investigate the inherent interactions in a multitask environment, especially focusing on the incremental value of invasiveness information

TABLE I
DISTRIBUTIONS OF PATIENTS IN THREE DATASETS

	Training (*EGFR*+/-)	Validation (*EGFR*+/-)	Testing (*EGFR*+/-)	Total (*EGFR*+/-)
Dataset 1	383 (205/177)	129 (69/60)	129 (69/60)	640 (343/297)
Dataset 2	/	/	50 (28/22)	50 (28/22)
Dataset 3	/	/	36 (9/27)	36 (9/27)

(*EGFR*+/-) shown as the number of cases with *EGFR*-mutant type vs *EGFR*-wiled type

for predicting EGFR mutation status from different aspects (see methods for detail). The finally designed original deep learning scheme, named GMILT, incorporated MIL, transformer, active learning, and multitask learning algorithms. MIL, which reduced the annotation requirements using coarse-grained input information, was conducive to solving the problem of uncertain positive areas in the lesion. In this article, attention-based MIL was used to purify and merge the effective feature information of each instance and then improve the model characterization ability. Moreover, the concept of pseudo-3-D was adopted in MIL. Drawing on the idea of active learning, we innovatively proposed the combined use of SAM and GAS loss by selecting valuable samples online for feature learning to improve the convergence speed and generalization ability of the deep learning model.

Considering that EGFR classification is related to the invasiveness characteristics of the lesion itself, we use the ensemble learning method to carry out multitask learning on visual words implemented by GEM to realize the structured coding of instance-level representation and further improve the representation ability.

C. GMILT Can Improve the Performance in Predicting EGFR Mutation Status

To mine highly discriminative features and improve the performance of EGFR prediction using CT images, we designed a multiple-instance learning transformer network (the baseline of GMILT). The baseline GMILT (model 2) achieved an AUC of 0.759 by using CT images only for predicting EGFR mutation status. To explore the interaction between the invasiveness information and EGFR mutation status, we designed three additional (models 3–5) by considering the invasiveness information as different supplemental information to the EGFR predictive network. Model 3, in which lung nodule mask and invasiveness information were both fed as the inputs, demonstrated a similar predictive efficiency for EGFR mutation status (model 3: AUC $=$ 0.721 versus model 2: AUC $=$ 0.759), indicating that the given information of invasiveness may be a confounding factor rather than a contributing factor. In model 4, inspired by the inherent link between invasiveness and EGFR mutation status, we considered both the invasiveness and EGFR mutation information as supervised factors and constructed a multitask model to simultaneously predict both. However, model 4 showed no improvement over model 2 in performance for predicting EGFR mutation status (AUC $=$ 0.700 versus model 2: 0.759). However, its performance for predicting invasiveness substantially improved relative to that of model 1 (AUC $=$ 0.926 versus model 1: 0.879). Finally, model 5 (GMILT) considered the invasiveness information as supplemental information embedded into the intermediate features, and the performance for predicting EGFR mutation status was improved, with the highest AUC of 0.772 in all models [see Table II and Fig. 6(a)]. Furthermore, the AUC of model 5 was significantly higher than that of model 4 ($P = 0.042$). It indicated that the invasiveness information may be related to the EGFR mutation status and can be appropriately used to improve the performance in predicting EGFR mutation status.

D. GMILT Obtained a Robust Performance on External and Public Testing Datasets and Excelled Over Other Methods

One of the limitations of data-dependent deep learning is the relatively weakened performance on external datasets due to the inconsistent data distribution. To validate the stability and generalizability of GMILT, we further tested the model on an independent external dataset and a public dataset. Our model obtained similar performance outcomes on the external dataset (AUC $=$ 0.756) and even better performance outcomes on the TCIA dataset (AUC $=$ 0.856) (see Table III) than on the internal testing dataset. Moreover, to investigate the efficiency of our proposed model, we selected three representative published papers and repeated their deep learning methods on our Dataset 1, constructed as model 2.5-D [8], model 3-D [10], and model SE-CNN [7]. Our proposed GMILT network excelled over the previous three methods (AUC: 0.772 versus 0.720, 0.741, and 0.649) [see Table III and Fig. 6(b)]. Together, these findings support the advantage of multi-instance learning and active learning in predicting EGFR mutation status. The precision curves and confusion matrix information are described in Sections 4–7 in the Supplementary Material.

E. GMILT Achieved Better Performance in Different Diameter-Based Subgroups

Inspired by the different tumor sizes observed in clinical scenarios, we further investigated the performance of **GMILT** on four size-based subgroups. As described in Table IV, most of the results were superior to those of model 5 (AUC $=$ 0.772) on the testing dataset of Dataset 1, obtaining state-of-the-art performance. GMILT obtained good performance in each subgroup, indicating that the constructed model had discriminative features for each sample data point (see Table IV).

F. Ablation Study of GMILT

The ablation study was typically used when testing a network by removing individual parts. To verify the effectiveness

TABLE II
PERFORMANCE OUR DESIGNED MODELS IN PREDICTING EGFR MUTATION STATUS

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUPRC	F1-score
Model1-IAC/non-IAC	0.879 (0.813,0.944)	0.829 (0.763,0.894)	0.75 (0.627,0.871)	0.877 (0.804,0.947)	0.783 (0.679,0.887)	0.855 (0.794,0.917)	0.842 (0.764,0.922)	0.766 (0.671,0.857)
Model2	0.759 (0.674,0.838)	0.744 (0.666,0.815)	0.812 (0.717,0.901)	0.667 (0.545,0.779)	0.737 (0.662,0.807)	0.755 (0.655,0.85)	0.753 (0.661,0.855)	0.772 (0.702,0.836)
Model3	0.721 (0.632,0.807)	0.713 (0.637,0.788)	0.754 (0.653,0.854)	0.667 (0.546,0.783)	0.722 (0.647,0.798)	0.702 (0.61,0.795)	0.736 (0.65,0.821)	0.738 (0.665,0.808)
Model4	0.700 (0.614,0.787)	0.659 (0.582,0.736)	0.522 (0.405,0.637)	0.817 (0.715,0.92)	0.766 (0.66,0.877)	0.598 (0.533,0.664)	0.754 (0.679,0.83)	0.621 (0.519,0.72)
Model4-IAC/non-IAC	0.926 (0.875,0.975)	0.899 (0.844,0.95)	0.854 (0.751,0.953)	0.926 (0.867,0.982)	0.872 (0.784,0.959)	0.915 (0.86,0.968)	0.902 (0.839,0.963)	0.863 (0.787,0.934)
GMILT	**0.772** (0.688,0.856)	0.752 (0.678,0.826)	0.667 (0.558,0.778)	0.85 (0.759,0.939)	0.836 (0.751,0.923)	0.689 (0.615,0.767)	0.803 (0.72,0.89)	0.742 (0.657,0.827)

-IAC/non-IAC shown as the performance in predicting IAC/non-IAC, others were presented the performance in predicting EGFR mutation status, 95% confidence interval was presented in parentheses. PPV=positive predictive value, NPV=negative predictive value, AUPRC=area under the curve of precision-recall curve.

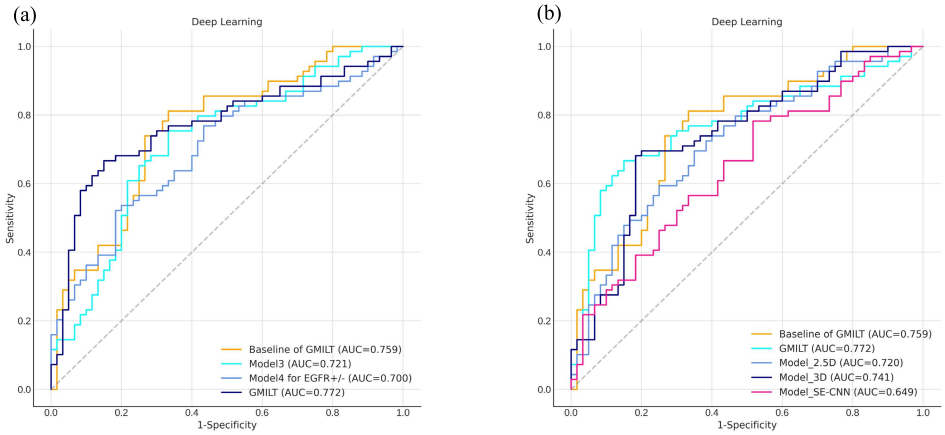


Fig. 6. Performance of 5 proposed models in (a) our study and (b) comparative results with 2.5-D, 3-D, and SE-CNN models (trained and validated on our dataset).

of our newly proposed parts in GMILT, we performed the ablation study on our two key modules (GEM and SAM+GAS). We investigated the effect of separately adding these two modules to GMILT on performance outcomes. Both modules outperformed the baseline of GMILT (AUC: 0.769 and 0.763 versus 0.759) (see Table V and Fig. 7), indicating the efficiency of the two modules in representation learning.

G. Visualization and Interpretation Analysis

Considering the heterogeneous nature of tumors, we tried to identify the core area from which the pathological characteristics were most likely to originate. The two attention pooling layers in our GMILT block, inner attention pooling and outer attention pooling, model the relationships among transaxial, coronal, and sagittal patches and visual words, respectively. We constructed attention maps of different queries in the transformer encoder to reveal the potential areas that contributed most to the predicted results (see Fig. 8). After 3-D transformation, the potential "core area" (red dot in Fig. 9) could be visualized and used to guide biopsy and minimize false negatives.

H. t-SNE Visualization of the Features Learned by GMILT

To intuitively explore the manifold structure of the features, we visualized the features by testing Dataset 1 using t-distributed stochastic neighbor embedding (t-SNE) [33], which is particularly suitable for the visualization of high-dimensional data. As illustrated in Fig. 10, the EGFR+ [yellow dots in Fig. 10(a)] and EGFR- [purple dots in Fig. 10(a)] samples formed two distinct sample clusters. Samples predicted to be EGFR+ were more likely to be located closer to the upper left corner, and those predicted as EGFR- were more likely to be located closer to the lower right corner [see Fig. 10(a)]. The prediction scores are consistent with the ground truth [see Fig. 10(b)]. Furthermore, our model was verified to be effective in representation learning and could construct highly

TABLE III
PERFORMANCE ON TESTING DATASETS AND COMPARISON RESULTS

Model and dataset	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUPRC	F1-score
Testing performance								
GMILT in external dataset (Dataset 2)	0.756 (0.615,0.906)	0.76 (0.648,0.881)	0.75 (0.596,0.909)	0.773 (0.61,0.95)	0.808 (0.695,0.938)	0.708 (0.577,0.856)	0.703 (0.562,0.897)	0.778 (0.666,0.895)
GMILT in TCIA dataset (Dataset 3)	0.856 (0.657,1.000)	0.917 (0.832,1.000)	0.778 (0.52,1.000)	0.963 (0.891,1.000)	0.875 (0.678,1.000)	0.929 (0.854,1.000)	0.834 (0.642,1.000)	0.824 (0.628,1.000)
Comparison analysis in internal testing dataset of Dataset 1								
Model-2.5D	0.720 (0.631,0.807)	0.674 (0.592,0.756)	0.696 (0.587,0.804)	0.65 (0.529,0.771)	0.696 (0.614,0.779)	0.65 (0.558,0.744)	0.734 (0.639,0.831)	0.696 (0.614,0.776)
Model-3D	0.741 (0.653,0.833)	0.744 (0.67,0.821)	0.681 (0.574,0.79)	0.817 (0.718,0.918)	0.81 (0.725,0.901)	0.69 (0.614,0.771)	0.756 (0.67,0.847)	0.74 (0.658,0.824)
Model-SE-CNN	0.649 (0.541,0.742)	0.643 (0.558,0.721)	0.783 (0.681,0.87)	0.483 (0.35,0.617)	0.635 (0.571,0.705)	0.659 (0.538,0.775)	0.682 (0.591,0.778)	0.701 (0.627,0.763)

95% confidence interval was presented in parentheses. PPV=positive predictive value, NPV=negative predictive value, AUPRC=area under the curve of precision-recall curve.

TABLE IV
DISTRIBUTIONS AND PERFORMANCE OF SUBGROUP ANALYSIS

	EGFR+	EGFR-	All (EGFR+/-)	AUC
	6[a]	10	16	0.583
0<MD<1cm	2[b]	1	3	**1.000**
	0[c]	1	1	0.718*
	22	31	53	0.740
1≤MD<2cm	0	3	3	**0.875**#
	3	12	15	0.718*
	17	5	22	**0.824**
2≤MD<3cm	7	2	9	**0.857**
	2	6	8	**1.000**
	24	14	38	0.756
3cm≤MD	19	16	35	**0.789**
	4	8	12	**0.875**

[a, b, c] represents the validation dataset 1, 2, and 3 respectively. * represents the AUC calculated for the subgroup 0<MD<2cm in dataset 3. # represents the AUC calculated for the subgroup 0<MD<2cm in dataset 2. Bold numbers indicate the performance is better than model 5 (0.772).

discriminative features, allowing for improved performance in predicting EGFR expression status.

IV. DISCUSSION

Noninvasively predicting the EGFR mutation status is a persistent challenge but represents an urgent need in the clinic. While deep learning has its own advantages in this area, its performance is limited by the need to learn efficient and discriminative features related to EGFR mutation status. Inspired by the potential inherent links between EGFR mutation status and invasiveness information, we hypothesized that the predictive performance of a deep learning network can be improved through extra utilization of the invasiveness information. Thus, in this study, we proposed a new deep learning model, **GMILT**, to predict EGFR mutation status. To the best of our knowledge, this is the first study to investigate the interaction effects between EGFR mutation status and invasiveness information, and it is also the first study to introduce the transformer method to medical tasks. Our study found that utilizing invasiveness information as embedding features in the network can substantially improve its performance. Our proposed model achieved an AUC of 0.772, with favorable generalizability to a public dataset and external validation dataset (AUC $=$ 0.856 and 0.756, respectively). In addition, the proposed model performed better for size-specific subgroups with state-of-the-art (65% $\sim$ 81% in previous studies) classification AUCs and, thus, can be applied in different clinical scenarios. Finally, our model can

TABLE V
PERFORMANCE OF THE ABLATION STUDY

	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	AUPRC	F1-score
GMILT	0.772 (0.688,0.856)	0.752 (0.678,0.826)	0.667 (0.558,0.778)	0.85 (0.759,0.939)	0.836 (0.751,0.923)	0.689 (0.615,0.767)	0.803 (0.72,0.89)	0.742 (0.657,0.827)
Baseline of GMILT (Model 2)	0.759 (0.674,0.838)	0.744 (0.666,0.815)	0.812 (0.717,0.901)	0.667 (0.545,0.779)	0.737 (0.662,0.807)	0.755 (0.655,0.85)	0.753 (0.661,0.855)	0.772 (0.702,0.836)
Baseline of GMILT+GEM	0.769 (0.687,0.851)	0.721 (0.641,0.8)	0.652 (0.534,0.768)	0.8 (0.7,0.9)	0.789 (0.7,0.881)	0.667 (0.586,0.749)	0.795 (0.712,0.881)	0.714 (0.621,0.804)
Baseline of GMILT+SAM+GAS	0.763 (0.677,0.847)	0.736 (0.661,0.809)	0.623 (0.505,0.74)	0.867 (0.779,0.951)	0.843 (0.754,0.932)	0.667 (0.595,0.74)	0.794 (0.71,0.879)	0.717 (0.622,0.807)

95% confidence interval was presented in parentheses. PPV=positive predictive value, NPV=negative predictive value, AUPRC=area under the curve of precision-recall curve.

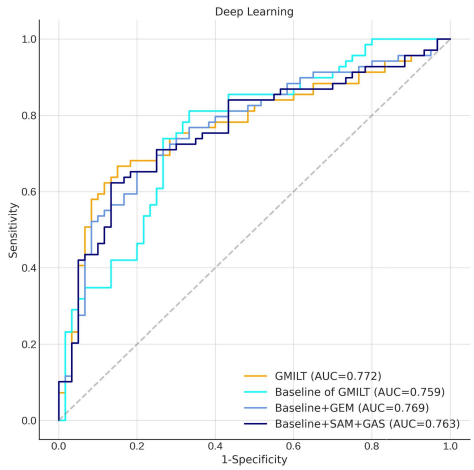


Fig. 7. ROC curves of three ablation studies.

visualize the potential "core area" that most correlates to EGFR expression to guide biopsy or pathological evaluation procedures.

In comparisons of data mining approaches in medical scenarios, deep learning seems to be superior to other methods (e.g., radiomics). This superiority has also been seen in the task of predicting EGFR mutation status [8], [10]. In this study, our findings demonstrated the same tendency (see Section 8 in the Supplementary Material). However, the current deep learning models in predicting EGFR mutation status could be further improved by efficiently and deeper mining the higher dimensional features or relationships Invasiveness information has been proven to have potential inherent links with EGFR mutation status [12], [13]. However, no previous study had investigated the incremental value of this link to deep learning models for predicting EGFR mutation status. In this context, we proposed a new network, GMILT, which can efficiently exploit and utilize all discriminative patterns (i.e., the relationship between EGFR mutation status and invasiveness) bridging the representation gap in the spatial and global information of the tumor to predict the EGFR mutation status.

Ideally, adding the label as an input into the network may improve the performance. However, the opposite result was obtained (AUC: $0.721 < 0.759$). This finding suggests that directly using a pathological label as an input cannot make the deep learning model efficiently mine its intrinsic relationship with EGFR typing. Inspired by the successful experience in our previous multitask study (i.e., applying the segmentation task to improve the classification task), we also investigated the potential mutual promotion of these factors in a multitask environment, considering the invasiveness information as supervised learning information instead of an input. The performance of predicting EGFR mutation status also failed to be improved, indicating that the features or links related to EGFR mutation status are more complicated and high level (AUC: $0.700 < 0.759$). In contrast, the performance of predicting invasives was improved. This may indicate that features related to EGFR mutation status might provide valuable supplemental information in predicting the invasiveness of lung adenocarcinoma. Meanwhile, features related to the invasiveness status were relatively more obvious and relevant, and could easily be learned by the model. Since pathological invasiveness information and the EGFR category belong to two completely different dimensions of information, the inherent correlation is unclear, which makes it difficult to effectively drive the model to use pathological information to directly promote EGFR classification. In this context, we used ensemble learning to structure and quantitatively construct the representation space of supervised learning in the study so that the network can effectively improve the efficiency of EGFR classification by facilitating auxiliary tasks in control. This strategy improved the performance of the model in predicting EGFR mutation status (AUC: $0.772 > 0.759$).

Although the improvement was slight, it is difficult to make a breakthrough in predicting EGFR mutational status using CT images. This may be partly attributed to the reason that the features or correlates regarding gene status are more comprehensive and difficult to learn than those correlated with other tasks, such as the prediction of benign or malignant tumors [34], the prediction of multiple pathological types [35], and the risk stratification of lung adenocarcinoma [36]. Several clinical factors, such as smoking and sex, are well-known factors related to EGFR mutations [37]. Adding these clinical factors can improve predictive performance. Note that our proposed model presented a more effective result than clinical

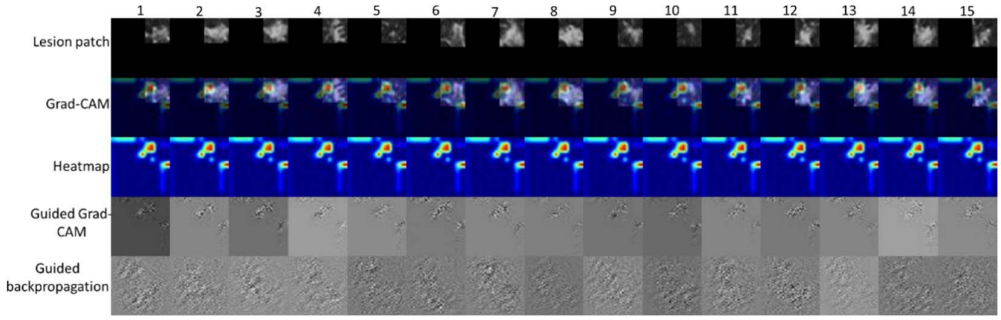


Fig. 8. Suspicious areas (heat map) of EGFR mutant generated by Grad-CAM network within 15 patches from transaxial, coronal, and sagittal levels.

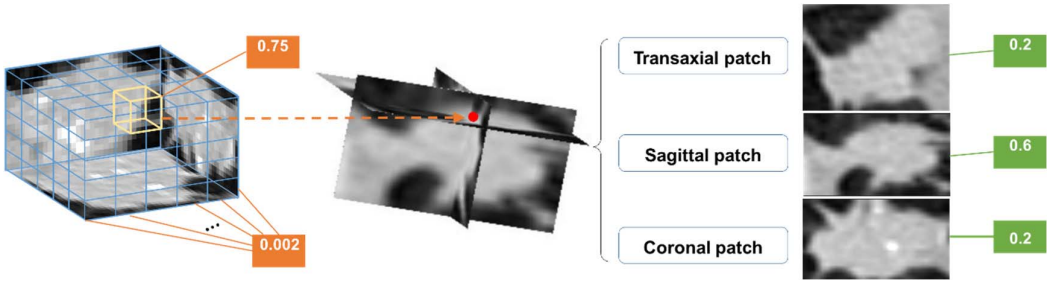


Fig. 9. Visualization of the attention mechanism in our method. The attention weights of visual words and transaxial, coronal, and sagittal patches for a lesion are presented above as orange and green boxes. The orange box shows the highest weight value of 0.75, which indicates the most probable EGFR mutant area in the lesion. 0.2, 0.6, and 0.2 represent the probability of EGFR mutant in the transaxial, sagittal, and coronal patches, respectively.

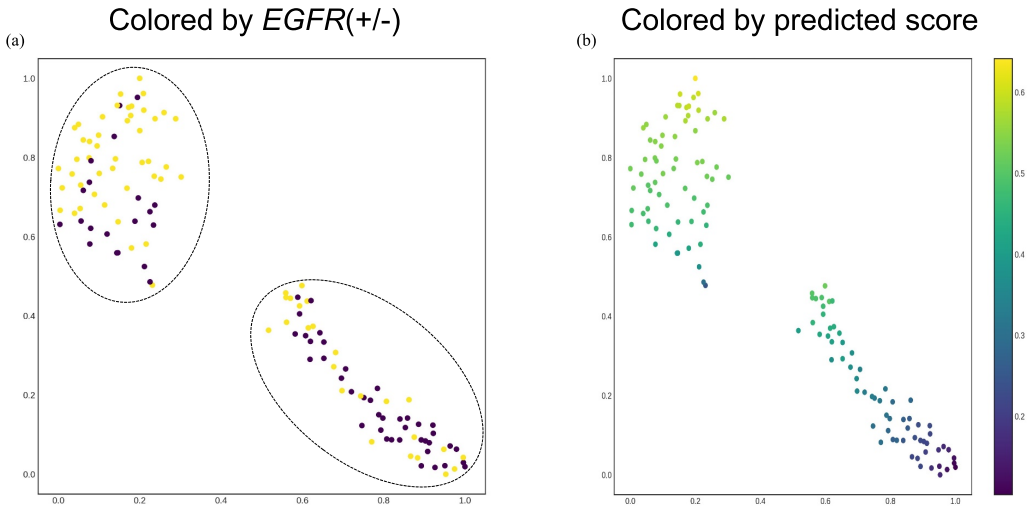


Fig. 10. Visualization of features in our internal test dataset using the t-SNE technique in a 2-D space. (a) Yellow and purple represent EGFR+ and EGFR- samples, respectively. (b) t-SNE visualization scatter plot colored by the EGFR probability score predicted by ***GMILT***. The prediction scores are consistent with the ground truth.

models [38] or radiomics combined with clinical factor models [39], [40], only using CT images. Moreover, our proposed model also outperformed traditional machine learning methods (i.e., radiomics) and previous deep learning methods in comparative analysis (AUC: 0.772 versus 0.720, 0.741, and 0.649). This indicates that our model may facilitate high-level feature learning by embedding invasive information into the network to fully utilize the potential correlations or patterns hidden behind the tumor. Images from other modalities, e.g., PET-CT, have also recently been used to predict EGFR mutation status with deep learning [41]. The authors found that models fed fusion images (i.e., PET/CT images) outperformed the CT image-based model. However, the data sample size was relatively small, and the performance of the CT image-based model (AUC = 0.72) was inferior to ours. Moreover, their network SE-ResNets neglects the interactive information between the two layers, resulting in an inefficient and unnecessary way to channel attention learning [42].

The performance of our model was substantially improved in the subgroup analysis (most of the AUCs were over 0.772). The current results also illustrated the effectiveness of our model in the process of feature learning and the improvement of the feature discrimination ability. Generally, the larger the lesion is, the more heterogeneous it will be, leading to a more complex feature distribution. In this case, using our model to construct the discriminant characterization of the sample and then subgroup prediction could achieve higher predictive efficacy.

In the clinic, false negatives may lead to an inappropriate treatment regimen, thus compromising the prognosis. False negatives are possibly attributed to intratumor heterogeneity [43], [44]. In this context, predicting the core area most likely to be related to the true EGFR mutation status in advance can substantially minimize false negatives. Motivated by this clinical need, we used multiple attention layers in our model to visualize the importance of lesion areas affecting EGFR typing and improved the interpretability of the model. It is worth noting that our proposed model uses a nested hierarchical structure to realize the uncoupling of feature learning and abstraction processes, which can not only obtain the probability of lesionwise cubes but also further obtain the probability of the three axial planes corresponding to each cube. This strategy is similar to the decision tree. Therefore, an alternative way of interpretability is innovatively proposed in this study, which can locate the significant area of the lesion volume. With this guidance provided by the model, clinicians can better find the area expressing EGFR and excise tissue cores.

In general, images contain a lot of background information (i.e., noise) and limited effective or valuable information. In another word, images in the real world are generally low-quality data or low-SNR data. CNNs are generally prone to noise interruptions, i.e., small image noise can cause drastic changes in the output and lead to overfitting [45]. In this study, we adopted MIL and transformer to reduce the noise effect and improve the SNR of the input information. This strategy has been proven to resolve low-quality data [46]. Based on previous MIL, we innovatively built pseudo-3-D instances to ensure that each instance had enough spatial and semantic information. The transformer structure is designed to improve the discrimination of characterization information and efficiently use spatial and global information to improve the SNR in the representation space. Considering that EGFR classification is related to the invasiveness characteristics of the lesion itself, we used the ensemble learning method to carry out multitask learning on visual words implemented by GEM to realize the structured coding of instance-level representation. At the same time, the organic combination of SAM and GEM promotes the effectiveness of online sample selection, thus improving the performance (AUC: model 5 $>$ model 2). Our proposed method provides a new approach for analyzing medical data that could be considered either a reference or method for investigators performing other tasks (i.e., predicting prognosis or treatment of cancers).

There were still several conceived limitations in our study. First, this was a retrospective study and only verified it in the current task. A further prospectively designed study and application in other tasks are warranted to verify the efficiency of the model. Second, the sample size was relatively small, and the data distribution may be unbalanced and biased due to the limited number of centers that participated in our study. Larger sample size and more participating centers can lead to better performance. However, despite such a small dataset, we yielded comparable and promising results compared with other studies, especially in subgroup analysis. The newly proposed technique for analyzing medical data introduced in our study can provide a novel methodology for other medical tasks. Finally, although attention pooling can present the potential core area to help clinicians in the biopsy, it is at the "proof-of-concept" stage. Slice-level comparisons with gross tissues (for pathological analysis) are needed to confirm the efficiency of this technique. However, our model uses an attention mechanism to define the weight of the "core area," which can substantially improve the accuracy of key area identification [17], [47].

V. Conclusion

In this article, we proposed a novel network-GMILT to noninvasively predict EGFR mutation status, which can be applied in different clinical scenarios regarding patients with lung adenocarcinoma. Moreover, the visualization analysis shows the ability of our model to reveal the potential "core area" that most correlates to EGFR expression and, thus, facilitate the application of precision medicine.

References

[1] R. L. Siegel *et al.*, "Cancer statistics, 2020," *CA Cancer J. Clin.*, vol. 70, no. 1, pp. 7–30, 2020.
[2] D. S. Ettinger *et al.*, "NCCN guidelines insights: Non–small cell lung cancer, version 1.2020," *J. Nat. Comprehensive Cancer Netw.*, vol. 17, no. 12, pp. 1464–1472, 2019.
[3] V. A. Miller *et al.*, "Molecular characteristics of bronchioloalveolar carcinoma and adenocarcinoma, bronchioloalveolar carcinoma subtype, predict response to erlotinib," *J. Clin. Oncol.*, vol. 26, no. 9, pp. 1472–1478, Mar. 2008.
[4] T. S. Mok *et al.*, "Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma," *New England J. Med.*, vol. 361, no. 10, pp. 947–957, 2009.
[5] K. Hastings *et al.*, "*EGFR* mutation subtypes and response to immune checkpoint blockade treatment in non-small-cell lung cancer," *Ann. Oncol.*, vol. 30, no. 8, pp. 1311–1320, 2019.
[6] H. Bai *et al.*, "Influence of chemotherapy on *EGFR* mutation status among patients with non–small-cell lung cancer," *J. Clin. Oncol.*, vol. 30, no. 25, pp. 3077–3083, 2012.
[7] B. Zhang *et al.*, "Deep CNN model using CT radiomics feature mapping recognizes EGFR gene mutation status of lung adenocarcinoma," *Frontiers Oncol.*, vol. 10, Feb. 2021, Art. no. 598721.
[8] S. Wang *et al.*, "Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning," *Eur. Respiratory J.*, vol. 53, no. 3, Mar. 2019, Art. no. 1800986.
[9] Y. Dong *et al.*, "Multi-channel multi-task deep learning for predicting EGFR and Kras mutations of non-small cell lung cancer on CT images," *Quant. Imag. Med. Surg.*, vol. 11, no. 6, pp. 2354–2375, Jun. 2021.
[10] W. Zhao *et al.*, "Toward automatic prediction of EGFR mutation status in pulmonary adenocarcinoma with 3D deep learning," *Cancer Med.*, vol. 8, no. 7, pp. 3532–3543, Jul. 2019.
[11] S. Moreno *et al.*, "A radiogenomics ensemble to predict EGFR and Kras mutations in NSCLC," *Tomography*, vol. 7, no. 2, pp. 154–168, Apr. 2021.
[12] S. B. Yoo, J.-H. Chung, H. J. Lee, C.-T. Lee, S. Jheon, and S. W. Sung, "Epidermal growth factor receptor mutation and p53 overexpression during the multistage progression of small adenocarcinoma of the lung," *J. Thoracic Oncol.*, vol. 5, no. 7, pp. 964–969, Jul. 2010.
[13] H. Chen *et al.*, "Genomic and immune profiling of pre-invasive lung adenocarcinoma," *Nature Commun.*, vol. 10, no. 1, p. 5472, 2019.
[14] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3614–3633, Jul. 2021.
[15] K. Han *et al.*, "Transformer in transformer," 2021, *arXiv:2103.00112*.
[16] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[17] Z. Li *et al.*, "A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101978.

[18] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Proc. MICCA*, 2020.
[19] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101789.
[20] T. Vu, P. Lai, R. Raich, A. Pham, X. Z. Fern, and U. A. Rao, "A novel attribute-based symmetric multiple instance learning for histopathological image analysis," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3125–3136, Oct. 2020.
[21] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102062.
[22] K. Clark *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013.
[23] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.
[24] H. Chen and A. Shrivastava, "Group ensemble: Learning an ensemble of ConvNets in a single ConvNet," 2020, *arXiv:2007.00649*.
[25] Y. Wen *et al.*, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
[26] W. L. Ranger. A synergistic optimizer. GitHub. [Online]. Available: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer
[27] L. Liu *et al.*, "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*.
[28] M. R. Zhang, J. Lucas, J. Ba, and G. E. Hinton, "Lookahead optimizer: *k* steps forward, 1 step back," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–12.
[29] H. Yong, J. Huang, X. Hua, and L. Zhang, "Gradient centralization: A new optimization technique for deep neural networks," 2020, *arXiv:2004.01461*.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[31] G. E. Hinton *et al.*, "Improving neural networks by preventing co-adaptation of feature detectors," *Comput. Sci.*, vol. 3, no. 4, pp. 212–223, 2012.
[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.
[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2015.
[34] D. Ardila *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Med.*, vol. 25, no. 6, pp. 954–961, Jun. 2019.
[35] Y. Fu *et al.*, "Fusion of 3D lung CT and serum biomarkers for diagnosis of multiple pathological types on pulmonary nodules," *Comput. Methods Programs Biomed.*, vol. 210, Oct. 2021, Art. no. 106381.
[36] J. Gong *et al.*, "Deep learning-based stage-wise risk stratification for early lung adenocarcinoma in CT images: A multi-center study," *Cancers*, vol. 13, no. 13, p. 3300, Jun. 2021.
[37] G. Singal *et al.*, "Association of patient characteristics and tumor genomics with clinical outcomes among patients with non–small cell lung cancer using a clinicogenomic database," *J. Amer. Med. Assoc.*, vol. 321, no. 14, pp. 1391–1399, 2019.
[38] J. Zhao *et al.*, "CT characteristics in pulmonary adenocarcinoma with epidermal growth factor receptor mutation," *PLoS ONE*, vol. 12, no. 9, Sep. 2017, Art. no. e0182741.
[39] E. R. Velazquez *et al.*, "Somatic mutations drive distinct imaging phenotypes in lung cancer," *Cancer Res.*, vol. 77, no. 14, pp. 3922–3930, Jul. 2017.
[40] Y. Liu *et al.*, "Radiomic features are associated with EGFR mutation status in lung adenocarcinomas," *Clin. Lung Cancer*, vol. 17, no. 5, pp. 441–448.e6, Sep. 2016.
[41] G. Yin *et al.*, "Prediction of EGFR mutation status based on 18F-FDG PET/CT imaging using deep learning-based model in lung adenocarcinoma," *Frontiers Oncol.*, vol. 11, Jul. 2021, Art. no. 709137.
[42] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13–19.
[43] T. Zhang *et al.*, "Genomic and evolutionary classification of lung cancer in never smokers," *Nature Genet.*, vol. 53, no. 9, pp. 1348–1359, 2021.
[44] H. Zhou *et al.*, "Multi-region exome sequencing reveals the intratumoral heterogeneity of surgically resected small cell lung cancer," *Nature Commun.*, vol. 12, no. 1, Dec. 2021, Art. no. 5431.
[45] Q. Li *et al.*, "Wavelet integrated CNNs for noise-robust image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7245–7254.
[46] Z. Wang, J. Poon, S. Wang, S. Sun, and S. Poon, "A novel method for clinical risk prediction with low-quality data," *Artif. Intell. Med.*, vol. 114, Apr. 2021, Art. no. 102052.
[47] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. Ö. Arik, and T. Pfister, "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3417–3425.

Wei Zhao received the Ph.D. degree in imaging and nuclear medicine from Fudan University, Shanghai, China, in 2019.

He is currently an Associate Professor with Central South University, Changsha, China, where he is also the Assistant Director of the Radiology Department, The Second Xiangya Hospital. His research interests include imaging, radiomics, and deep learning.

Dr. Zhao is also the Young Talent in Hunan Province and received the Hunan Provincial Natural Science Foundation for Excellent Young Scholars.

Weidao Chen received the master's degree in biomedical engineering from Zhejiang University, Hangzhou, China, in 2018.

He is currently a Machine Learning Researcher with InferVision Medical Technology Company Ltd., Beijing, China. His research interests include medical image algorithms and medical artificial intelligence.

Ge Li received the M.S. degree in imaging and nuclear medicine from Central South University, Changsha, China, in 2016.

She is currently a Radiology Technologist with the Xiangya Hospital, Central South University. Her research interests include medical imaging and breast imaging.

Du Lei received the Ph.D. degree in radio physics from East China Normal University, Shanghai, China, in 2011.

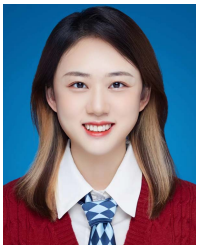
He is currently a Research Scientist with the College of Medicine, University of Cincinnati, Cincinnati, OH, USA. His research interests include magnetic resonance imaging (MRI) and medical artificial intelligence.

Jiancheng Yang (Member, IEEE) received the B.Eng. and M.Eng. degrees in automation from Shanghai Jiao Tong University, Shanghai, China, in 2015 and 2018, respectively, and the Engineer Degree (double master's degree) from the Institut Mines-Télécom, Évry-Courcouronnes, France, in 2016. He is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University.

He was a Visiting Researcher with Harvard University, Cambridge, MA, USA, and the École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland. His research interests center around the interdisciplinary field of medical image analysis and 3-D computer vision.

Yanjing Chen received the bachelor's degree in medical imaging from Fujian Medical University, Fuzhou, China, in 2019. She is currently pursuing the master's degree in radiology with the Xiangya Hospital, Central South University, Changsha, China.

She is currently a Research Assistant with The Second Xiangya Hospital, Central South University.



Yingjia Jiang received the bachelor's degree in medical imaging from Shanxi Medical University, Jinzhong, China, in 2019.

She is currently a Research Assistant with The Second Xiangya Hospital, Central South University, Changsha, China. Her research interests include neuroimaging and diagnosis.

Jiangfen Wu received the Biomedical Engineering degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2014.

She is currently in charge of the Institute of Translational Medicine, InferVision Medical Technology Company Ltd., Beijing, China. Her research interests include medical image algorithms and medical artificial intelligence.

Bingbing Ni received the B.Eng. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2005, and the Ph.D. degree from the National University of Singapore, Singapore, in 2011.

He was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He was a Research Intern with Microsoft Research Asia, Beijing, China, in 2009. He was a Software Engineer Intern with Google Inc., Mountain View, CA, USA, in 2010. He is currently a Professor with the Department of Electrical Engineering, Shanghai Jiao Tong University.

Dr. Ni was a recipient of the First Prize in the International Contest on Human Activity Recognition and Localization in conjunction with the International Conference on Pattern Recognition in 2012.

Yeqi Sun received the Ph.D. degree in clinical pathology from the School of Medicine, Shanghai Jiao Tong University, Shanghai, China, in 2019.

She is currently an Attending Pathologist with The Second Xiangya Hospital, Central South University, Changsha, China. Her research interests include molecular pathology and diagnosis.

Shaokang Wang received the master's degree in statistics from The University of Chicago, Chicago, IL, USA, in 2012.

He is currently in charge of the application of artificial intelligence technology in medical imaging at InferVision Medical Technology Company Ltd., Beijing, China. His research interests include medical image algorithms and medical artificial intelligence.

Yingli Sun received the M.D. degree in imaging and nuclear medicine from Fudan University, Shanghai, China, in 2018.

She is currently an Attending Radiologist with the Huadong Hospital Affiliated to Fudan University. Her research interests include lung disease diagnosis and machine learning.

Ming Li received the Ph.D. degree in imaging and nuclear medicine from Fudan University, Shanghai, China, in 2012.

He is currently a Chief Physician with the Huadong Hospital Affiliated to Fudan University. His research interests include medical imaging and medical artificial intelligence.

Jun Liu is currently a Professor with Central South University, Changsha, China, where he is also the Director of the Radiology Department, The Second Xiangya Hospital. His research interests include brain functional imaging, radiomics, and deep learning.

Dr. Liu is also a National Member of the Neurology Group of the Chinese Society of Radiology, National Committee of the Neurology Group of the Radiological Branch of the Chinese Medical Association. He is also the Technological Leading Talent and the Leader of 225 subjects in Hunan Province.