Feature Copy-Paste Network for Lung Cancer EGFR Mutation Status Prediction in CT images

Xingyu Huang, Shuo Wang *, Chengcai Liu, Haolin Sang, Yi Wu, and Jie Tian

Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing, China

Abstract. Epidermal growth factor receptor (EGFR) mutation status is crucial for targeted therapy planning in lung cancer. Current identification relies on invasive biopsy and expensive gene sequencing. Recent studies indicate that CT imaging with advanced deep learning techniques offer a non-invasive alternative for predicting EGFR mutation status. However, CT scanning parameters, such as slice thickness, vary significantly between different scanners and centers, making the predicting models highly sensitive to data types, and thus not robust in clinical practice. In this study, we propose Feature Copy-Paste Network (FCP-Net), an innovative and robust model for predicting EGFR mutation status using CT images. First, we propose a novel Feature Copy-Paste Consistency (FCPC) module to exchange the information from CT scans with different slice thicknesses and impose consistency constrain to make model more robust. Second, we introduce a Feature Refinement (FR) module to filter redundant features during information fusion, thereby enhancing the accuracy of mutation prediction. Extensive experiments demonstrate the outstanding performance of the FCPC and FR modules. When the trained model is tested on both thin-slice and thick-slice CT images, it achieves at least 2.6% and 2.1% improvements in AUC, respectively, indicating the models' robustness and stability. Our code is available at `https://github.com/499huangxingyu/FCPNet`.

Keywords: EGFR gene mutation · Feature copy-paste · CT image.

1 Introduction

Epidermal growth factor receptor (EGFR) gene mutation status is critical for the treatment planning of lung cancer[5]. In clinical practice, EGFR mutation status is determined by biopsy and gene detection[9],[4], which is invasive, expensive and can cause false-positive results due to tumor genetic heterogeneity[16]. Computed Tomography (CT) is a non-invasive and effective technique widely used in lung cancer analysis. The easy accessibility of CT scans, along with their potential correlations with gene expression patterns, enhances their utility and enables CT to serve as a valuable tool for both clinical applications and research[12,2].

* Corresponding author: shuo_wang@buaa.edu.cn

In earlier studies, radiomics-based methods have been explored for predicting EGFR gene mutations[14],[15]. While these methods rely on handcrafted features and often suffer from limited generalizability across datasets due to imaging variability and feature instability. To overcome these limitations, deep learning has emerged as a promising method that automatically learns hierarchical image representations and has shown strong potential in predicting EGFR gene mutations from CT imaging[17],[20],[19],[3]. Nevertheless, CT images are heterogeneous between different scanners and centers, which affects deep learning models largely. For instance, variations in slice thickness lead to differences in data distribution, which in turn impact the robustness of the model. To improve model robustness, Wang et al.[16] employed domain adaptation to learn the relationship between thin- and thick-slice images. This method simultaneously extracts and integrates information from both slice thicknesses, effectively alleviating heterogeneity between thin- and thick-slice data. Liu et al.[11] used contrastive learning to enhance the model's focus on lung parenchyma regions. By integrating global and local features, this method alleviated the impact of data heterogeneity and achieved robust performance in EGFR mutation status predicting. Although the aforementioned methods enhance model stability, the interaction between image features with varying slice thicknesses is crucial. If the model fails to adequately extract and interact with these features, it may become biased toward extracting information from a specific slice thickness, leading to an inability to generalize and resulting in unstable performance.

To fully enable interaction between different data groups and improve the network's robustness, Image Copy-Paste (ICP)[18],[1],[10],[6] is a simple yet powerful data augmentation technique widely used in computer vision tasks. It has the potential to encourage the model to learn the distribution and semantic information from diverse data groups. Based on ICP, utilizing copy-paste to exchange information between thin- and thick-slice images enables the model to focus on features that are consistent between two types of images. This approach has the potential to further improve the model's generalization capability. However, medical images, especially lung CT scans, place more emphasis on texture and anatomical details rather than overall contours, and this is where ICP has its limitations. Medical images require fine-grained attention to these details, and performing copy-paste operations at the image level may disrupt these critical features, leading to instability in model performance. Additionally, although ICP operations enhance information interaction and data diversity, they inevitably distort the original image information during training[7], increasing the difficulty of model prediction. In contrast, we believe that performing copy-paste operations in the feature space while enforcing consistency in both image and feature spaces can alleviate image heterogeneity and improve model stability.

In order to exchange data information more effectively and enhance model robustness, this paper proposes a novel Feature Copy-Paste (FCP) operation based on ICP and introduces a Feature Copy-Paste Consistency (FCPC) module to optimize the network. The FCP operation enables interaction between different data in the feature space, enhancing data diversity and connectivity,

while the FCPC module constrains various feature types to alleviate heterogeneity between data. Furthermore, to eliminate redundant features, we introduce Feature Refinement (FR) module to refine features and improve prediction accuracy. Extensive experiments demonstrate that the proposed method achieves promising performance on both thin- and thick-slice CT images and also performs well with limited training data, validating the model's robustness.

2 Methodology

Figure 1 illustrates the architecture of the proposed FCPNet. We randomly pick up a pair of CT scans (A, B) indicating thin- and thick-slice CT images of a patient as original inputs. Then, we extract a random patch PA' from image A and PB' from image B, and exchange PA' and PB' to generate the Image Copy-Paste (ICP) inputs (A', B'). Subsequently, the two input pairs are fed into a shared-weight feature encoder (DenseNet) to obtain features $(f_A, f_B, f_{A'}, f_{B'})$. These features are sequentially optimized through the FCPC and FR modules. Finally, the refined feature are utilized to predict lung cancer EGFR gene mutation status.

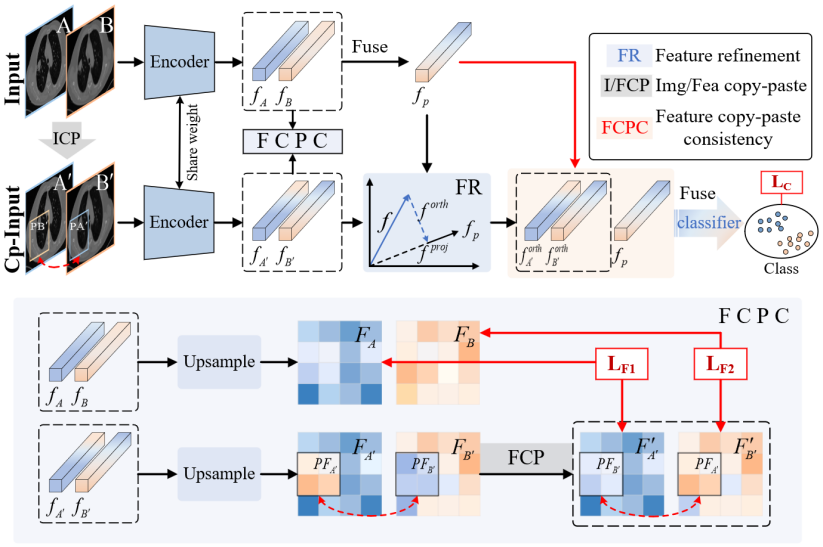

Fig. 1. Overview of FCPNet

2.1 Feature Copy-Paste Consistency Module

(Image Copy-Paste) ICP has been demonstrated to be an effective data augmentation strategy; However, while increasing data diversity, it may potentially loss

original features and reduce the model's robustness. To improve the model's stability and robustness, we introduce the Feature Copy-Paste Consistency (FCPC) module as shown in Figure 1. The two input image pairs are processed by the encoder to generate two corresponding feature sets: the original input features (f_A, f_B) and the ICP image features $(f_{A'}, f_{B'})$. These one-dimensional features are subsequently upsampled to match the size of the original images, obtaining (F_A, F_B) and $(F_{A'}, F_{B'})$. Similar to the ICP operation, patches $(PF_{A'}, PF_{B'})$ are extracted from the same locations in $(F_{A'}, F_{B'})$ and undergo a copy-paste operation to generate $(F'_{A'}, F'_{B'})$. The original images are processed through ICP to extract image features $(f_{A'}, f_{B'})$, followed by the FCP operation at the same locations to obtain corresponding features $(F'_{A'}, F'_{B'})$. Theoretically, these features should be consistent with the features of the original images (F_B, F_A). Therefore, feature consistency constraints are imposed between $(F'_{A'}, F_B)$ and $(F'_{B'}, F_A)$ to ensure that the network can enhance its stability and robustness when adapting to varying images.

2.2 Feature Refinement Module

In the network, four different one-dimensional features $(f_A, f_B, f_{A'}, f_{B'})$ are extracted through the encoder. However, these features include substantial redundancy. To prevent such redundancy, we introduce the Feature Refinement (FR) module. As demonstrated in figure 1, to better preserve the original image features, the network firstly integrates f_A and f_B from the primary feature f_p:

$$f_p = \mathrm{Conv}(\mathrm{C}(f_A, f_B)) \tag{1}$$

Where $\mathrm{Conv}(\cdot)$ and $\mathrm{C}(\cdot)$ represent the convolution layer and concatenation operation.

Subsequently, we use f_p as the template, and apply orthogonal projection to filter out redundant features from $f_{A'}$ and $f_{B'}$, as follows:

$$f_x^{orth} = f_x - (\frac{f_x \cdot f_p}{|f_p|^2}) f_p \quad x \in \{A', B'\} \tag{2}$$

Here, f^{orth} represents the refined features. Finally, a convolution layer is used to fuse f_x^{orth},$x \in \{A', B'\}$ and f_p to predict EGFR gene mutation status.

2.3 Loss Functions

Based on the above description, there are two optimization objectives in the model: the prediction loss and the feature copy-paste consistency loss. The prediction loss is computed using cross-entropy loss, while the feature copy-paste consistency loss is enforced by the mutual information (MI) loss, which can be expressed as:

$$
\begin{aligned}
L_F &= \mathrm{MI}(F_A, F'_{A'}) + \mathrm{MI}(F_B, F'_{B'}) \\
\mathrm{MI}(F_A, F'_{A'}) &= \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(F_A = {F_A}^{(i)}, {F'}_{A'} = {F'}_{A'}^{(i)})}{p(F_A = {F_A}^{(i)}) \cdot p({F'}_{A'} = {F'}_{A'}^{(i)})} \\
\mathrm{MI}(F_B, F'_{B'}) &= \frac{1}{N}\sum_{i=1}^{N} \log \frac{p(F_B = {F_B}^{(i)}, {F'}_{B'} = {F'}_{B'}^{(i)})}{p(F_B = {F_B}^{(i)}) \cdot p({F'}_{B'} = {F'}_{B'}^{(i)})}
\end{aligned} \tag{3}
$$

Here, $p(x, y)$ represents the joint probability, and $p(x)$ and $p(y)$ denote the respective marginal probabilities. the total loss can be described as:

$$L = L_C + \lambda L_F \tag{4}$$

Where λ=0.5 is the hyper-parameter to trade off the contributions of the losses.

3 Experiments and Results

3.1 Dataset and Preprocessing

The dataset included 3433 patients with lung cancer from the West China Hospital, each with EGFR gene detection result and multiple CT image scans of different slice thicknesses. CT images with slice thicknesses between [0.5, 2.5] mm were categorized as thin CT images (thin thickness domain), while those with slice thicknesses between [3, 8] mm were thick CT images (thick thickness domain). Since a patient may have multiple thin CT scans and thick CT scans, we generated multiple thin-thick image pairs for each patient. Afterward, the dataset was randomly divided into a training set with 2058 patients (3186 thin-thick image pairs), a validation set with 342 patients (543 image pairs), and a testing set (1033 patients with 1663 thin-thick image pairs). We first used a publicly available U-Net model to segment the lung ROI in all 3D CT images. Then, we resized the lung ROI to (256, 256, 48) voxel size and normalized it to standardize the CT voxel intensity.

3.2 Implementation Details

To validate the effectiveness of the proposed FCPNet, we compared it with recently reported models, including DenseNet-thin (DN.-thin, using only thin-slice images as input), DenseNet-thick (DN.-thick, using only thick-slice images as input), DenseNet-co (DN.-co, using both thin- and thick-slice scans as input)[8], QSNet[16], and PLCHNet[11]. All models were implemented using PyTorch[13] and trained with the SGD optimizer on NVIDIA GeForce RTX 3090 GPUs. In all the comparison models, the default settings were used, except for the replacement of the input with thin- and thick-slice CT images in PLCHNet.

3.3 Comparison Results

We used the accuracy (Acc), sensitivity (Sen), specificity (Spe) and area under the ROC curve (AUC) to evaluate the prediction performance. As shown in Table 1, the proposed FCPNet outperformed other SOTA models. Specifically, our method achieved improved generalization and stability in images with varying slice thickness compared to traditional DenseNet models. Compared to the domain adaptation-based network QSNet, FCPNet achieved Acc/AUC improvement of 0.90%/4.99% in test set with thin-slice images, and 0.72%/2.86% in test set with thick-slice images. Compared with contrastive learning model (PLCHNet), FCPNet achieved Acc/AUC improvement of 2.42%/2.64% in test set with thin-slice images and 1.79%/2.17% in test set with thick-slice images.

Table 1. Quantitative performance of different methods.

Test set	Matrics	Methods					
		DN.-thin	DN.-thick	DN.-co	QSNet	PLCHNet	FCPNet
With thin-slice CT images	Acc	72.760	71.737	69.753	73.000	71.918	**73.662**
	Sen	0.533	0.509	0.350	**0.743**	0.577	0.494
	Spe	0.757	0.704	**0.828**	0.512	0.699	0.800
	AUC	0.700	0.670	0.662	0.701	0.717	**0.736**
With thick-slice CT images	Acc	72.699	73.108	73.301	74.564	73.782	**75.105**
	Sen	0.663	0.584	0.589	**0.732**	0.617	0.604
	Spe	0.646	0.702	0.737	0.620	0.711	**0.749**
	AUC	0.703	0.718	0.714	0.733	0.738	**0.754**

Figure 2 presented the ROC curves of different methods on test set with thin-slice images and thick-slice images. As depicted in the figure 2, traditional DenseNet-based methods showed significant performance fluctuations across different test sets, indicating model instability. In contrast, QSNet and PLCHNet mitigated these issues partially. Comparatively, our FCPNet achieved the highest AUC value across the test sets with different slice thickness, demonstrating the stability and robustness of the proposed FCPNet.

3.4 Ablation Studies

FCPC and FR Modules Table 2 showed the quantitative results of different strategy combinations. From the table, we observed that the baseline model only achieved AUC=0.68 and 0.71 in thin and thick CT test sets. Gradual incorporation of the proposed modules brought progressive improvements in model performance. In contrast, compared to the baseline model, when all modules were integrated, the model reached optimal accuracy, with the test AUC for thin-slice image and thick-slice image improving by 7.76% and 6.20% respectively.

Feature Copy-Paste Consistency Loss Figure 3(a) illustrated the performance comparison of different feature copy-paste consistency losses. As shown

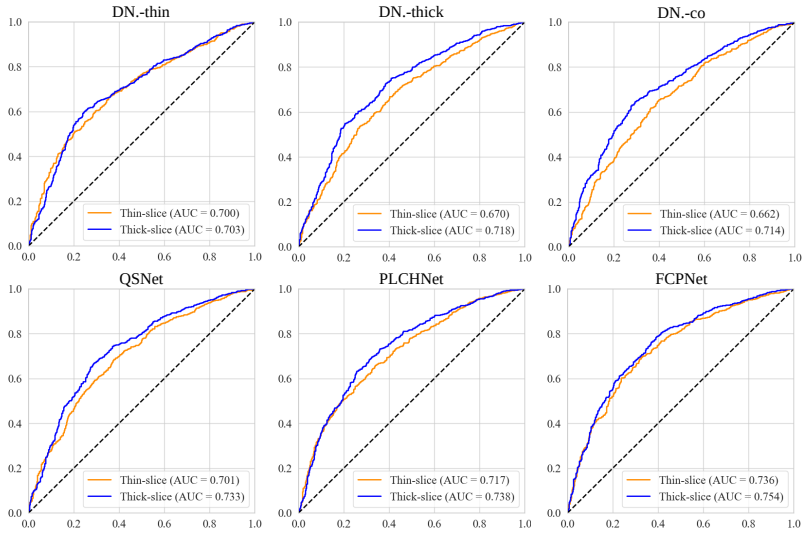

Fig. 2. ROC curves of different methods in the test set with thin-slice images and thick-slice images

Table 2. Effectiveness of feature copy-paste consistency strategy and relevant modules, with the baseline models using DenseNet. Bold values indicate the best result.

Methods	FCP	FLoss	FR	Test-thin CT		Test-thick CT	
				Acc	AUC	Acc	AUC
Baseline				72.519	0.683	73.060	0.710
	✓			72.339	0.658	73.361	0.706
	✓		✓	72.519	0.703	74.323	0.744
	✓	✓		72.760	0.699	72.820	0.724
	✓	✓	✓	**73.662**	**0.736**	**75.105**	**0.754**

in the figure 3(a), the model failed to achieve optimal performance without any copy-paste feature constraints (w/o_Floss). However, introducing feature consistency constraints largely improved prediction accuracy. Notably, when the loss function was set to MI, the model achieved optimal performance. Furthermore, introducing feature consistency constraints significantly improved prediction accuracy by better aligning feature distributions, with AUC values reaching 0.736 and 0.754 in the thin-slice and thick-slice test sets, respectively.

Patch Size of Copy-Paste Operations Figure 3(b) depicted the impact of using different patch sizes during the copy-paste operation. It was observed that neither excessively large nor small patches enabled the model to achieve optimal performance. We speculated that overly small patches failed to support adequate information exchange, while larger patches were unable to preserve the original

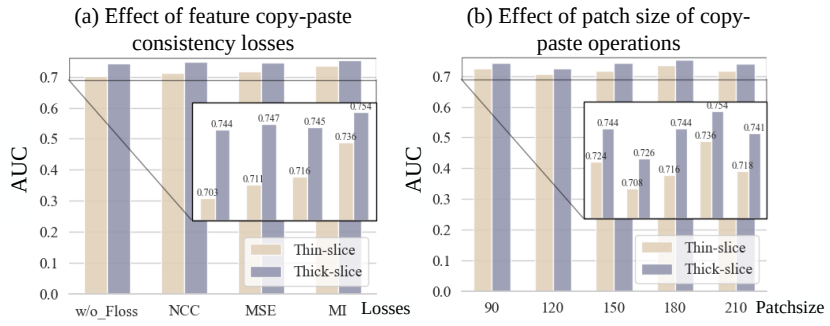

Fig. 3. (a) AUC values of different feature copy-paste consistency loss functions. (b) AUC values for various patch size in copy-paste operations

information. In our experiments, setting the patch size to (180, 180) allowed the model to achieve the best performance, with AUC values of 0.736 and 0.754 for thin-slice test set and thick-slice test set, respectively.

limited Training Data Figure 4 illustrated the AUC bar charts of different methods under various proportions of training data. We observed that DenseNet-based methods and QSNet were highly sensitive to the training data amount, with the model collapsing when the training set size was reduced to 30% or 20%. Although the performance degradation amplitude of the PLCHNet is relatively smaller than DenseNet models, its accuracy still decreased 7.1% when the training data amount is only 20%. In comparison, although the prediction accuracy of FCPNet also decreased with a reduction in training data, it remained relatively stable. Even with only 20% of the training data, the model achieved AUC values of 0.677 and 0.712 in thin-slice test set and thick-slice test set.

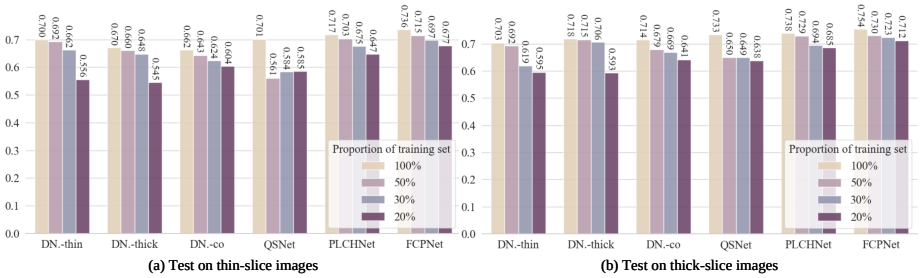

Fig. 4. AUC values of different methods under different training data amount. (a) AUC values tested in test set with thin-slice images, (b) AUC values in test set with thin-slice images

4 Conclusion

In this study, we proposed a novel Feature Copy-Paste Network (FCPNet) for predicting EGFR gene mutation status in lung cancer using CT images. The Feature Copy-Paste Consistency (FCPC) and Feature Refinement (FR) modules efficiently improved prediction performance. Notably, the model maintained superior performance even with limited training data, demonstrating its robustness and stability. Extensive experiments validated the effectiveness and superiority of our method, highlighting its potential clinical applicability.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (92259105), Beijing Natural Science Foundation (L232131), Ministry of Science and Technology of China (No.2023YFC2506502).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11514–11524. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01108
2. Burr, R., Leshchiner, I., Costantino, C.L., Blohmer, M., Sundaresan, T., Cha, J., Seeger, K., Guay, S., Danysh, B.P., Gore, I., Jacobs, R.A., Slowik, K., Utro, F., Rhrissorrakrai, K., Levovitz, C., Barth, J.L., Dubash, T., Chirn, B., Parida, L., Sequist, L.V., Lennerz, J.K., Mino-Kenudson, M., Maheswaran, S., Naxerova, K., Getz, G., Haber, D.A.: Developmental mosaicism underlying EGFR-mutant lung cancer presenting with multiple primary tumors. Nature Cancer **5**(11), 1681–1696 (Oct 2024). https://doi.org/10.1038/s43018-024-00840-y
3. Cheng, J., Liu, J., Jiang, M., Yue, H., Wu, L., Wang, J.: Prediction of Egfr Mutation Status in Lung Adenocarcinoma Using Multi-Source Feature Representations. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1350–1354. IEEE (2021). https://doi.org/10.1109/ICASSP39728.2021.9414064
4. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nature Medicine **24**(10), 1559–1567 (Oct 2018). https://doi.org/10.1038/s41591-018-0177-5
5. Fang, S., Wang, Z.: Egfr mutations as a prognostic and predictive marker in non-small-cell lung cancer. Drug design, development and therapy pp. 1595–1611 (2014)
6. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2917–2927. IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00294

7. Gong, C., Wang, D., Li, M., Chandra, V., Liu, Q.: KeepAugment: A Simple Information-Preserving Data Augmentation Approach. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1055–1064. IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00111
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. IEEE (2017). https://doi.org/10.1109/CVPR.2017.243
9. Jiang, Y., Ma, S., Xiao, W., Wang, J., Ding, Y., Zheng, Y., Sui, X.: Predicting EGFR gene mutation status in lung adenocarcinoma based on multifeature fusion. Biomedical Signal Processing and Control **84**, 104786. https://doi.org/10.1016/j.bspc.2023.104786
10. Kim, J., Jang, J., Seo, S., Jeong, J., Na, J., Kwak, N.: MUM: Mix Image Tiles and UnMix Feature Tiles for Semi-Supervised Object Detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14492–14501. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01411
11. Liu, M., Wang, S., Yu, H., Zhu, Y., Wang, L., Zhang, M., Wu, Z., Li, X., Li, W., Tian, J.: A Lung-Parenchyma-Contrast Hybrid Network For EGFR Gene Mutation Prediction In Lung Cancer. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE, Kolkata, India (Mar 2022). https://doi.org/10.1109/ISBI52829.2022.9761614
12. Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J., Aerts, H.J.W.L.: Foundation model for cancer imaging biomarkers. Nature Machine Intelligence **6**(3), 354–367 (Mar 2024). https://doi.org/10.1038/s42256-024-00807-9
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
14. Rossi, G., Barabino, E., Fedeli, A., Ficarra, G., Coco, S., Russo, A., Adamo, V., Buemi, F., Zullo, L., Dono, M., De Luca, G., Longo, L., Dal Bello, M.G., Tagliamento, M., Alama, A., Cittadini, G., Pronzato, P., Genova, C.: Radiomic Detection of EGFR Mutations in NSCLC. Cancer Research **81**(3), 724–731 (Feb 2021). https://doi.org/10.1158/0008-5472.CAN-20-0999
15. Tu, W., Sun, G., Fan, L., Wang, Y., Xia, Y., Guan, Y., Li, Q., Zhang, D., Liu, S., Li, Z.: Radiomics signature: A potential and incremental predictor for EGFR mutation status in NSCLC patients, comparison with CT morphology. Lung Cancer **132**, 28–35 (Jun 2019). https://doi.org/10.1016/j.lungcan.2019.03.025
16. Wang, L., Wang, S., Yu, H., Zhu, Y., Li, W., Tian, J.: A Quarter-split Domain-adaptive Network for EGFR Gene Mutation Prediction in Lung Cancer by Standardizing Heterogeneous CT image. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3646–3649. IEEE, Mexico (Nov 2021). https://doi.org/10.1109/EMBC46164.2021.9630395
17. Wang, S., Shi, J., Ye, Z., Dong, D., Yu, D., Zhou, M., Liu, Y., Gevaert, O., Wang, K., Zhu, Y., Zhou, H., Liu, Z., Tian, J.: Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. European Respiratory Journal **53**(3), 1800986 (Mar 2019). https://doi.org/10.1183/13993003.00986-2018
18. Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J.: CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In: 2019

IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6022–6031. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00612

19. Zhao, S., Li, W., Liu, Z., Pang, T., Yang, Y., Qiang, N., Zhao, J., Li, B., Lei, B., Han, J.: End-to-End Prediction of EGFR Mutation Status With Denseformer. IEEE Journal of Biomedical and Health Informatics **28**(1), 54–65 (2024). https://doi.org/10.1109/JBHI.2023.3307295
20. Zhao, W., Yang, J., Ni, B., Bi, D., Sun, Y., Xu, M., Zhu, X., Li, C., Jin, L., Gao, P., Wang, P., Hua, Y., Li, M.: Toward automatic prediction of *EGFR* mutation status in pulmonary adenocarcinoma with 3D deep learning. Cancer Medicine **8**(7), 3532–3543 (2019). https://doi.org/10.1002/cam4.2233