ResMLP_GGR: Residual Multilayer Perceptrons-Based Genotype-Guided Recurrence Prediction of Non-small Cell Lung Cancer

Yang Ai[1], Yinhao Li[1], Yen-Wei Chen[1,*], Panyanat Aonpong[2], and Xianhua Han[3]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan;
Email: gr0528fh@ed.ritsumei.ac.jp (Y.A.), yin-li@fc.ritsumei.ac.jp (Y.L.)
[2] Department of Computing, Faculty of Science, Silpakorn University, Nakorn Pathom, Thailand;
Email: aonpong_p@su.ac.th (P.A.)
[3] Artificial Intelligence Research Center, Yamaguchi University, Yamaguchi, Japan;
Email: hanxhua@yamaguchi-u.ac.jp (X.H.)
*Correspondence: chen@is.ritsumei.ac.jp (Y.W.C.)

***Abstract*—Non-small Cell Lung Cancer (NSCLC) is one of the malignant tumors with the highest morbidity and mortality. The postoperative recurrence rate in patients with NSCLC is high, which directly endangers the lives of patients. In recent years, many studies have used Computed Tomography (CT) images to predict NSCLC recurrence. Although this approach is inexpensive, it has low prediction accuracy. Gene expression data can achieve high accuracy. However, gene acquisition is expensive and invasive, and cannot meet the recurrence prediction requirements of all patients. In this study, a low-cost, high-accuracy residual multilayer perceptrons-based genotype-guided recurrence (ResMLP_GGR) prediction method is proposed that uses a gene estimation model to guide recurrence prediction. First, a gene estimation model is proposed to construct a mapping function of mixed features (handcrafted and deep features) and gene data to estimate the genetic information of tumor heterogeneity. Then, from gene estimation data obtained using a regression model, representations related to recurrence are learned to realize NSCLC recurrence prediction. In the testing phase, NSCLC recurrence prediction can be achieved with only CT images. The experimental results show that the proposed method has few parameters, strong generalization ability, and is suitable for small datasets. Compared with state-of-the-art methods, the proposed method significantly improves recurrence prediction accuracy by 3.39% with only 1% of parameters.**

***Keywords*—deep learning, non-small cell lung cancer, prediction of recurrence, radiogenomics, residual multilayer perceptrons**

I. Introduction

Lung cancer is the leading cause of cancer-related death worldwide. Non-small Cell Lung Cancer (NSCLC) includes all lung epithelial cancers, excluding small cell lung cancer, which is the main type of lung cancer, accounting for approximately 85% [1]. Thoracic surgeons prefer surgical resection for treating NSCLC. However, patients with NSCLC still have a high risk of recurrence and death after surgery and adjuvant chemotherapy. Once relapsed, more than half of the patients survived less than one year [2]. Accurately predicting the recurrence of NSCLC can effectively guide the clinical formulation of personalized treatment and follow-up plans, assist clinical decision-making, and improve the patients' survivability.

Currently, most studies on the prediction of cancer recurrence use radiomics methods [3, 4]. Taking X-ray, computed tomography (CT), magnetic resonance, ultrasound, nuclear medicine, and other medical images as research objects, researchers have attempted to use machine learning methods to extract high throughput quantitative features that can quantify tumor characteristics from medical images, and further mine on this basis to construct prediction models to achieve non-invasive analysis of tumor heterogeneity. The association between radiomics and cancer recurrence has been demonstrated for NSCLC [5–8]. Wang *et al.* [6] used the optimized features of multiple CT images and patient prognosis data as inputs, and applied classifiers such as random forest and Principal Component Analysis (PCA) to assist doctors in planning postoperative treatment and review of patients with NSCLC. To predict the recurrence of lung cancer within two years, Lee *et al.* first applied the Relief-F algorithm to select three significant features from 68 radiomics features, and then using random forest and support vector machine (SVM), they trained two groups of data representing recurrence and non-recurrence, respectively, within two years [7]. In a recent study, Christie *et al.* [8] predicted the recurrence of NSCLC by combining quantitative and qualitative features of the tumor and its surrounding area with clinical features. All studies have demonstrated that handcrafted features are potential biomarkers for predicting NSCLC recurrence.

Manuscript received December 1, 2022; revised March 20, 2023; accepted April 30, 2023.

With the development of Deep Learning (DL) technology, many DL methods based on CT images have also been applied to the classification of malignant tumors [9, 10]. Jiao *et al.* [11] developed a DL model combining circulating tumor cell counts and deep learning radiomics to predict the recurrence outcomes of patients with early-stage (ES) NSCLC treated with stereotactic body radiation therapy. Fusing pathological images of hematoxylin and eosin and clinical data, Yang *et al.* [12] developed a novel multimodal DL model that can accurately predict and evaluate the risk of recurrence and metastasis in HER2-positive breast cancer. With the powerful learning ability of DL, these methods can automatically quantify and extract deeper, clinically meaningful high-level features from medical images, without requiring too much manual intervention (pre-defined), and have outperformed radiomics methods. High-level features, largely abstracted by DL methods, can be used to predict NSCLC recurrence in an unbiased manner.

In recent years, some researchers have started exploring the potential of genetic information in prediction tasks. The researches [13, 14] proposed an approach that fused radiomics and genomics to predict lung cancer recurrence, achieving good results. Methods utilizing genetic information of genes achieve better prediction accuracy than machine learning and DL methods. However, the acquisition of gene expression is an invasive procedure and is significantly more expensive than the acquisition of CT images. In many cases, the required gene expression is undetectable and cannot be obtained in inoperable patients. The limitations of the research [13] are the possibility of surgery, the unavailability of gene expression data, and the high cost due to the complexity of the examination, which is an invasive diagnostic method.

Due to the association between images and gene expression [15–18], in our previous studies, we proposed a Genotype-Guided Radiomics (GGR) method and a deep genotype-guided radiomics fusion (GGR_fusion) method [17, 18], both of which were composed of a gene estimation model and a recurrence prediction model. However, the design of the gene estimation models is too simple to effectively use the features of different levels and is limited in their ability to comprehensively represent the potential information related to recurrence. Moreover, the estimates for each gene are separated, which complicates the training process. Besides, this dimension design, where 2,000 neurons are directly reduced to two neurons is unstable and may affect prediction accuracy. Hence, we proposed a novel residual multilayer perceptrons-based genotype-guided recurrence (ResMLP_GGR) prediction of NSCLC. First, for the gene estimation model, a ResMLP module is designed to replace the Fully Connected (FC) layer to fully integrate features of different modalities and dimensions (handcrafted and deep features), while avoiding the problem of gradient vanishing caused by a very deep structure. Furthermore, considering the stability of the recurrence prediction model, while keeping the width unchanged, a Deeper Multilayer Perceptron (MLP) model is developed to learn more complex expressions of the estimated genes to improve the accuracy of predicting recurrence. The main contributions of this study are as follows:

(1) We proposed a new feature extraction and fusion unit, the ResMLP module. This module is primarily composed of two MLPs with residual structures in series, which can fully fuse features of different modalities and dimensions.

(2) We developed a novel end-to-end gene estimation model that significantly reduces the number of parameters and increases the processing speed. In a small sample task, a model with fewer parameters avoids overfitting and improves its generalization ability. The end-to-end structure increases the number of learning samples in the encoder, further improving the generalization ability of the model.

(3) Experimental results demonstrate that the proposed method achieves state-of-the-art prediction accuracy compared to other radiomics and DL methods. Moreover, in the testing phase, the proposed algorithm can accurately predict the recurrence of NSCLC only by using non-invasive CT images.

Preliminary work was presented as a 4-pages conference paper in the 2022 IEEE Engineering in Medicine and Biology Society (EMBC) [19]. This paper involves methodological and experimental extensions and validations. We evolve the gene estimation model into an end-to-end structure to estimate 74 genes at one time, reducing the number of model parameters, speeding up the speed and improving the accuracy. In addition, we add ablation experiments and expand the comparison of the Fragments Per Kilobase Million (FPKM) values of genes estimated with the existing GGR method [17], the proposed method and the real genes to demonstrate the effectiveness of our proposed method.

II. Materials

This study used a publicly available radiogenomics dataset of NSCLC from the Open Research Cancer Imaging Archive (TCIA), funded by the National Cancer Institute [20]. The dataset provides CT images of NSCLC subjects (in Dicom format) and gene data from resected tumor tissue samples. The R01 cohort consisted of 162 patients with early-stage NSCLC, including 38 women and 124 men, with a mean scan age of 68 years (range: 42–86 years). They all gave written consent in accordance with the guidelines of institutions' IRBs [21].

A. Image Data and Preprocessing

At the request of the institution, all 162 subjects' CT image data were obtained in the supine position, with a single breath-hold scan ranging from the apex of the lung to the adrenal gland. According to the NSCLC recurrence task, the inclusion criteria for data used in this study were as follows: (a) patients with paired data (CT image and gene data); (b) patients with tumor mask and recurrence label. The exclusion criteria in this study were as follows: (a) patients without tumor masks; (b) patients without gene data; (c) patients without recurrence labels; (d) patients who died for unknown reasons prior to recurrence [18–21]. Finally, 88 patients who met the experimental

requirements were included in this study, and their clinical characteristics are listed in Table I.

For the segmentation of tumors, the initial segmentation of axial CT image sequences of the subjects was performed using an unpublished automatic segmentation algorithm [21]. These slides were then examined by radiologists with more than five years of experience diagnosing chest lesions, and the ePAD software was used to perform the necessary editing and corrections [21]. Finally, to evaluate the consistency of tumor segmentation among the participants, another radiologist with superior experience in diagnosing thoracic diseases reviewed the segmentation results of the previous doctors, and discussed and revised the dissenting samples until a consistent final segmentation result was obtained [21]. All ultimately approved segmentations (tumor masks) are stored as Dicom segmentation objects. Some CT images and their tumor masks are depicted in Fig. 1.

TABLE I. CLINICAL CHARACTERISTICS OF THE 88 PATIENTS

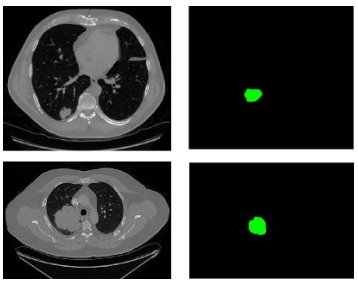
		Total n = 88 patients
Age (year)		46–85
Gender	Male Female	64 (72.72%) 24 (27.27%)
Smoking Status	Current Former Nonsmoker	14 (15.91%) 59 (67.05%) 15 (17.05%)
Recurrence	No Yes	59 (67.05%) 29 (32.95%)

Because the information contained in CT images is relatively complex, directly analyzing the original CT images will introduce a lot of noise unrelated to recurrence. To achieve accurate recurrence prediction of NSCLC, some preprocessing of CT images is required [22–24]. Considering the differences in tumor size between the subjects, the number of slices containing tumors generally varies. We first selected the largest slice of the tumor and its adjacent slices (the above and below slices, three slices for each patient) as the research objects and truncated the intensity information outside the range of −1000HU ~ +400HU (Hounsfield Unit). This range covers all the information required in the CT images in this study, including air, water, and soft tissues of different lungs [23]. Using the slice with the largest tumor area and its adjacent slices as input results in more volume data information and can better learn features related to the recurrence task than just select the largest slice alone [24], which is also a commonly used processing method in medical image preprocessing. The largest slice and its adjacent slices can be easily fed to the ResNet-50 model to extract deep features. These three slices of each sample were then normalized to pixel values ranging from 0 to 255 to mask intensity differences between different acquisition devices. Finally, we multiplied three CT images by their corresponding tumor masks, cropped the area of the bounding box of the CT images covered by the mask images, and resized the cropped image to 224 × 224 pixels by bicubic interpolation. The preprocessed CT images (ROIs) will be used for subsequent selection and extraction of handcrafted and deep features.

B. Gene Data and Selection

The gene data were obtained from tumor samples that surgeons remove from untreated subjects during surgery. The specific operation is as follows: cut a section 3-5mm thick slice along the longest axis of excised tumor tissue for RNA sequencing.

The samples were analyzed using RNA test data based on the availability and quality of the tissue collected. The gene expression for each patient was 22,127 genes [21], which were estimated by fragments per kilobase of transcript per million (FPKM) units [20–22]. However, most patients have absent gene expression, or N/A. Finally, 5587 genes with definite gene expression shared by each patient were initially screened and included as research objects for the gene data.



(a) CT images (b) tumor masks

Figure 1. Two examples of CT image data: (a) CT images (b) tumor masks.

Because the genes are still very large after initial screening, and the purpose of this study is to predict the recurrence of NSCLC, we only need to focus on genes related to recurrence, and genes unrelated to recurrence should be eliminated. Based on this, three feature selection methods (Chi-squared (CHI-2), F-test (analysis of variance; ANOVA) and least absolute shrinkage and selection operator (LASSO)) were used to select recurrence-related genes before developing the regression model, and their intersection was taken to determine the genes used for estimation [25]. Ultimately, 74 genes were screened. Please refer to Section III for more details.

III. METHOD

As shown in Fig. 2, the proposed method includes four steps: image preprocessing, feature extraction and selection, gene estimation and recurrence prediction. Different from conventional radiomics and DL methods that use a single prediction model, the proposed method is a dual design of a gene estimation model and a recurrence prediction model, based on the concept of GGR and GGR_fusion [17, 18], to maximizing model prediction accuracy and performance. The gene estimation model innovatively connected multiple ResMLP modules in series to learn the mapping relationship between CT images and genes by using handcrafted features, deep features and paired gene data, and then, the prediction model is used to complete the high-precision recurrence

prediction task using the estimated gene expression. In the testing phase, only CT images are required to complete the prediction, without any gene data, which is conducive to all patients with NSCLC and allows for the examination of patients who cannot undergo surgery.

Next, we introduce the proposed algorithm in detail. As image preprocessing has already been introduced, we will focus on feature extraction and selection, gene estimation, and recurrence prediction next.

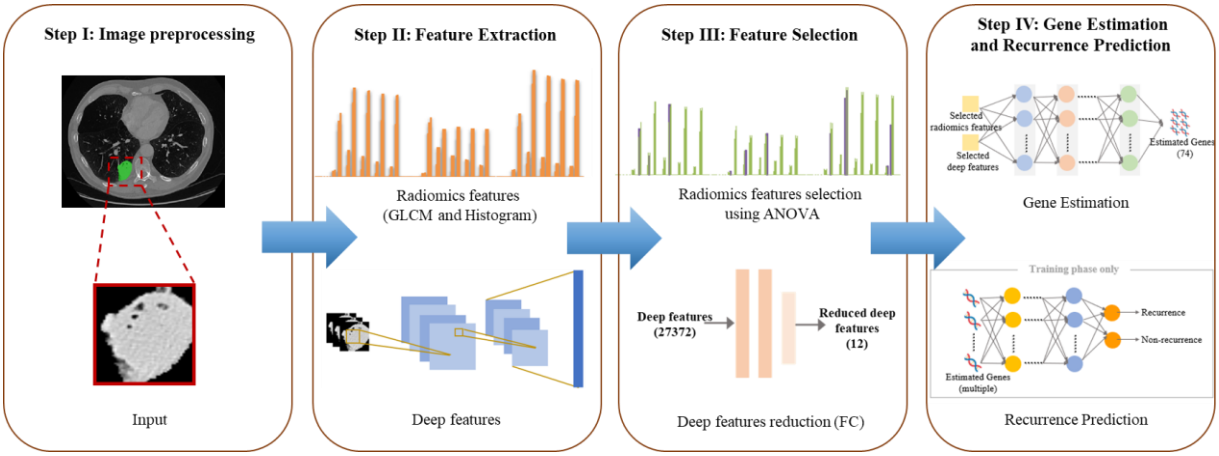


Figure 2. Overview of residual multilayer perceptrons-based genotype-guided recurrence (ResMLP_GGR) model prediction of non-small cell lung cancer.

A. *Feature Extraction and Selection*

To perform better gene estimation and use gene estimation information to guide and improve prediction accuracy, this study first developed a gene estimation model. This model combines hand-extracted low-level features with DL-extracted high-level semantic features to enrich feature sources, which serve as inputs to the gene estimation model. For each subject, feature extraction and selection included CT image feature (handcrafted and deep features) extraction and selection and genetic data extraction and selection [26–28].

1) *Handcrafted feature extraction*

To extract and select handcrafted features, we use the preprocessed ROIs from the previous section, i.e., three slices per patient, with each slice serving as a three-channel input. First, low-level features are extracted for each slice, and statistical features such as Standard Deviation (SD), mean, SD percentile, mean percentile, skewness, and kurtosis of CT images are extracted using the histogram statistics method [27]. The mean and SD percentiles were taken as 10, 25, and 50 percentiles, respectively. In this study, the texture features of CT images are also extracted. Ohanian et al. proposed that the Gray Level Co-occurrence Matrix (GLCM) outperformed several other commonly used methods for extracting texture features [28]. Therefore, we use the GLCM to extract texture features.

Suppose $f(x, y)$ is a two-dimensional digital image with a size of $M \times N$, the maximum gray level of the image is k, and the GLCM is the probability $P(i, j)$ of the occurrence of pixel (x, y) with a gray level of i and pixel $(x + a, y + b)$ with a gray level of j in the statistical image.

$$P(i,j) = \{(x, y), (x + a, y + b) \in M \times N \mid (f(x, y) = i, \\ f(x + a, y + b) = j)\}, \quad (1)$$

where P denotes the matrix $k \times k$, the pixel value of the image (x, y) is i, and the pixel value of the image $(x + a, y + b)$ is j. If the distance between (x, y) and $(x + a, y + b)$ is d, and the angle with the x-axis is θ, then the GLCM $P(i, j, d, \theta)$ related to d and θ is obtained. θ indicates the generation direction of the GLCM, and it is typically selected as 0°, 45°, 90° and 135°.

The GLCM represents the characteristics of the gray value of the image in any direction of space, but it cannot directly measure the difference between the textures of the target images. Haralick *et al.* [29] constructed 14 types of statistics based on the GLCM as texture features, but if all the 14 types of statistics are extracted on the basis of the GLCM, the calculation cost will be very high, which is undesirable in practical situations. The study by Ulaby et al. [30] proved that energy, contrast, correlation, entropy and homogeneity are uncorrelated, and their classification accuracy is high. The proposed method used the GLCM to extract energy, contrast, correlation, entropy, and homogeneity (inverse disparity), and other texture features in each slice at 0°, 45°, 90° and 135° directions (four-way GLCM).

In this study, the Laplacian of Gaussian (LoG) with five different filter parameters (σ = 0, 1, 1.5, 2, and 2.5) were used when extracting features by from ROI images using histogram statistics and GLCM methods [9, 17–18, 27]. LoG combines Gaussian and Laplacian filters that smooth an image to reduce noise and focus on things like rapidly changing areas or edge detection. Fig. 3 depicts the process of handcrafted feature extraction. 150 handcrafted features were extracted from each slice. After that, we used ANOVA to select 12 relevant handcrafted features [9] as one of the inputs to the gene estimation model. Some studies have shown that F-test [26, 27] can achieve the highest accuracy and Area Under Curve (AUC) among various feature selection methods when performing regression prediction using CT images. Hence, in this study, we used the F-test to

select relevant radiomics features for predicting recurrence-related genes, i.e., for estimating gene expression. As shown on the left (pink color) of Fig. 4, after F-test, 12 handcrafted radiomics features were selected to participate in gene estimation.

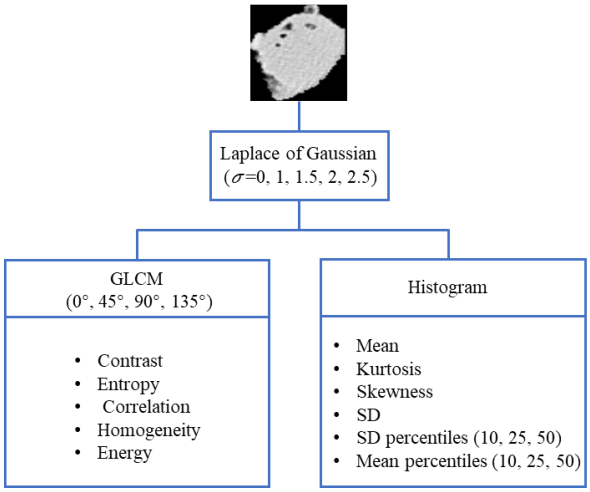


Figure 3. The process of extracting handcrafted features from each CT image band.

2) Deep feature extraction

Compared with the subjective visual assessment of images by clinically trained doctors, DL can automatically identify features in data and provide quantitative assessments [9, 10]. To better reflect the heterogeneity information of tumors and comprehensively characterize the characteristics of CT images related to NSCLC recurrence, we extracted both low-level handcrafted radiomics features and deep semantic information by ResNet-50 to better estimate gene expression [31, 32]. We initialize ResNet-50 with pretrained weights from the ImageNet dataset and freeze the FC layers for deep feature extraction. He et al. proposed ResNet, which is composed of many residual modules. ResNet-50 is a residual network with 50 layers (the network depth is 50). ResNet-50 is frequently used for various tasks [31]. As shown in blue on the left of Fig. 4, using the preprocessed ROI image as input, F-test is added between the last convolutional layer and the FC layer of the pretrained ResNet-50 network for feature screening. Applying the F-test yielded the same number of deep features (12 features) as handcrafted features most associated with NSCLC recurrence [26, 27], which was another input to the gene estimation model.

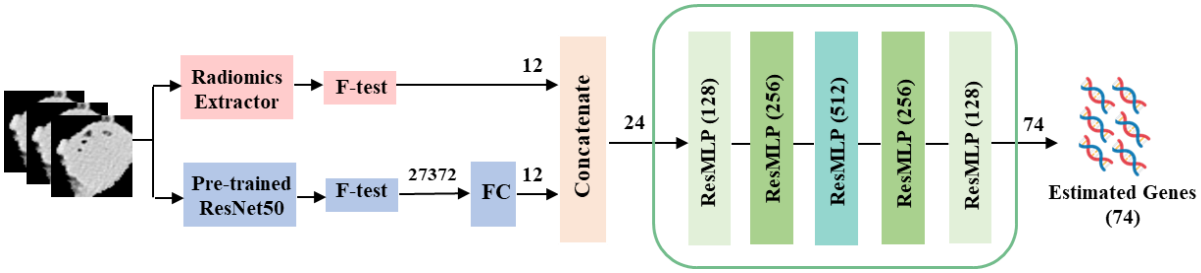


Figure 4. Structure of gene estimation regression model. The black numbers on the line show the features' number that passes the process.

B. Gene Estimation

To use gene expression information to improve the accuracy of the NSCLC recurrence prediction task, we first developed a gene estimation model, which used CT images to learn the expression of gene data, and then applied the learned gene estimation information to recurrence prediction.

At present, a deep neural network is typically used to develop regression models. However, Orhan *et al.* proved through experiments that with an increase in the network depth, the singular values of the product matrix become increasingly concentrated, whereas a few singular values with very low frequency become arbitrarily large [33]. These singularities produce degenerate manifolds on the surface of the loss function, reducing the efficiency of learning. Skipping connections will remove these singularities and speed up learning. Considering the small number of training samples, if a deep model or complex model is used, it will, result in overfitting and poor performance and poor generalization ability. Inspired by Subramanian *et al.* [33], this study developed the ResMLP module as the feature extraction and fusion unit of the gene estimation regression model.

As shown in Fig. 5, the proposed ResMLP module is composed of two MLPs with residual structures in series. The shortcut connections perform identity mapping, and their outputs are added to the outputs of the stacked layers to extract and fuse feature information of different levels, modalities and dimensions, which can comprehensively represent and fit potential gene expression functions related to NSCLC recurrence. For any input, it is first fed into the dense layer (x neurons), nonlinear activation layer (ReLU), and dropout layer. After the dropout, the output was put in the shortcut of the residual structure MLP. Identity mapping was performed by skip connection, and its output was added to the output of the stacked layer, passing through the two residual structure MLPs in turn. Each residual MLP consists of a Batch Normalization (BN) layer, a dense layer with x neurons, a ReLU layer, and a dropout layer. The BN layer can control the variance of the output of the previous layer, making the model more stable and increasing the training speed. The BN layer seeks to find a solution to the problem in a smoother solution subspace, as smoother solutions are more generalizable. The skip connection causes the model to have a more flexible structure [33]. When training, the model can adapt to its structure. Dropout emphasizes robustness, which

requires solutions to be insensitive to perturbations in the network configuration. In other words, a more robust solution has better generalization ability.

TABLE II. THE RESULTS FROM DIFFERENT GENE SELECTION METHODS

Feature selection method	Selected genes	Accuracy (%)
Non-selected	5587	81.14
LASSO	1123	82.97
F-test	131	86.89
CHI-2	2325	83.39
Intersection of LASSO, F-test, and CHI-2 (p-value < 0.02)	**74**	**94.30**

In this study, the estimation of genes comprehensively uses information from different modalities, including both hand-extracted prior knowledge (12 handcrafted features) and high-level semantic information (12 deep features) learned and screened by a deep network (ResNet-50). We concatenated these 24 relapse-related multimodal representations to estimate gene expression (regression). The estimated genes used the selected 74 genes in Section II.

The gene estimation model is depicted in Fig. 4. The gene estimation model consists of five ResMLP modules, with Arabic numbers in brackets representing the number of neurons. To increase the model's processing speed, we use an end-to-end structure for the gene estimation model. Thus, the estimation of 74 genes no longer requires several regression models of the same structure with different weight parameters. An estimate of all 74 genes was achieved using only one regression model can be considered an ensemble model assembled from a set of paths, where different paths contain different subsets of network layers. Here, we set the learning rate to 2e-4 to fit the gene estimation model. The optimizer uses NAdam, and the loss function is the mean squared error, which can be calculated by Eq. (2).

$$L_{mse} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad (2)$$

where L_{mse} denotes the mean square error loss, n denotes the number of data points, y denotes the actual gene label, and $\hat{y}$ denotes the gene prediction.

The gene estimation model with an end-to-end structure (uses only one gene estimation model) achieves faster convergence and is more conducive to optimizing and improving the generalization ability.

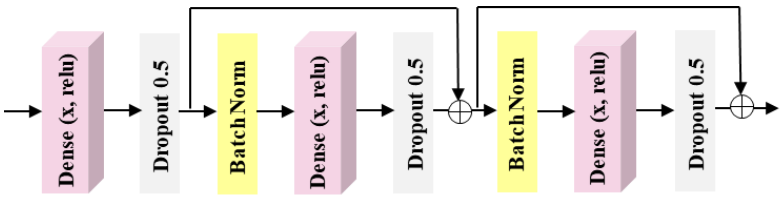


Figure 5. Residual Multilayer Perceptrons (ResMLP) module.

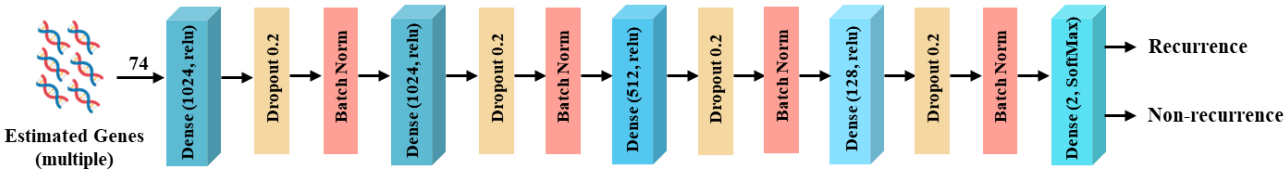


Figure 6. Structure of recurrence prediction model.

C. Recurrence Prediction

The proposed NSCLC recurrence prediction model is a robust binary classification model, which uses the 74 genes obtained by the previously proposed gene estimation model as input training a classification network, and learns richer and more advanced features by deepening the layers of the network. Compared with GGR and GGR_fusion, using fewer neurons results in higher robust classification accuracy on small datasets. In the testing phase, accurate NSCLC recurrence prediction can be achieved using only CT images without additional data. Here, the learning rate is set to 8e–4, and the loss function uses binary cross-entropy. The details of the proposed NSCLC recurrence prediction model are illustrated in Fig. 6.

The loss function of the NSCLC prediction model is defined as:

$$L_{bce} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln \hat{y}_i + (1-y_i)\ln(1-\hat{y}_i)), \quad (3)$$

where L_{bce} denotes the binary cross-entropy loss, n denotes the number of data, y denotes the actual recurrence label, and $\hat{y}$ denotes the recurrence prediction output.

IV. EXPERIMENT

All experiments in this study were conducted on a graphics accelerator-driven personal computer equipped with an RTX2080 Super 8GB discrete graphics card, an Intel (R) Core (TM) i7-10870H CPU @ 2.20GHz-2.21 GHz CPU and 16GB random access memory. We used python 3.7 and version 2.2.0 of the TensorFlow-GPU library as the experimental environment. As mentioned in Sections II and III, our experiment selected the data of 88 patients from the TCIA radiogenomics dataset of NSCLC for recurrence prediction [20]. The data of each patient

include three ROI CT images and 74 different gene data. Each experiment was repeated 10 times with 10-fold cross-validation to obtain a fair average performance. In each fold, CT images and gene expression data of patients (training set size of 79 or 80) were used in the training phase, whereas CT images of patients (validation set size of 9 or 8) were used only in the validation phase for recurrence prediction without gene expression data. We conducted an ablation experiment and compared related methods with ours with four different evaluation metrics using the same conditionally controlled dataset, and visualized the comparison results utilizing the Receiver Operating Characteristic (ROC) curve to demonstrate the superiority of our proposed method.

A. Gene Selection Results

As we mentioned in Section II, we applied three different feature selection methods and their intersection to select the genes most associated with NSCLC recurrence. As presented in Table II, LASSO removed gene features with zero coefficients [27], and the thresholds of CHI-2 and F-test were set to p-value < 0.02, respectively [25–27]. The first row in the table is the baseline, with a total of 5587 genes, and the last row is the 74 genes obtained by the intersection of the above three methods. According to the results in the table, the prediction accuracy of NSCLC recurrence is the highest when the intersection of the three methods is used. Therefore, the 74 genes obtained from the intersection of these three methods were selected for our regression model.

B. Gene Estimation Results

We fully used handcrafted and deep features as input to the gene estimation model to learn from real gene data (ground truth) and obtain gene estimation information as shown in Fig. 7. As the complexity of the gene data, we calculated the mean FPKM values of the estimated gene expression and the real gene data for all patients. As shown in the figure, the horizontal axis lists the names of 74 genes of a patient in and the vertical axis indicates the FPKM value of each gene. The blue line in the figure represents the real gene, the orange line represents the estimated gene obtained using the GGR method, and the gray line represents the estimated gene expression obtained using our proposed ResMLP_GGR method. Most of the estimated genes exhibited satisfactory performance.

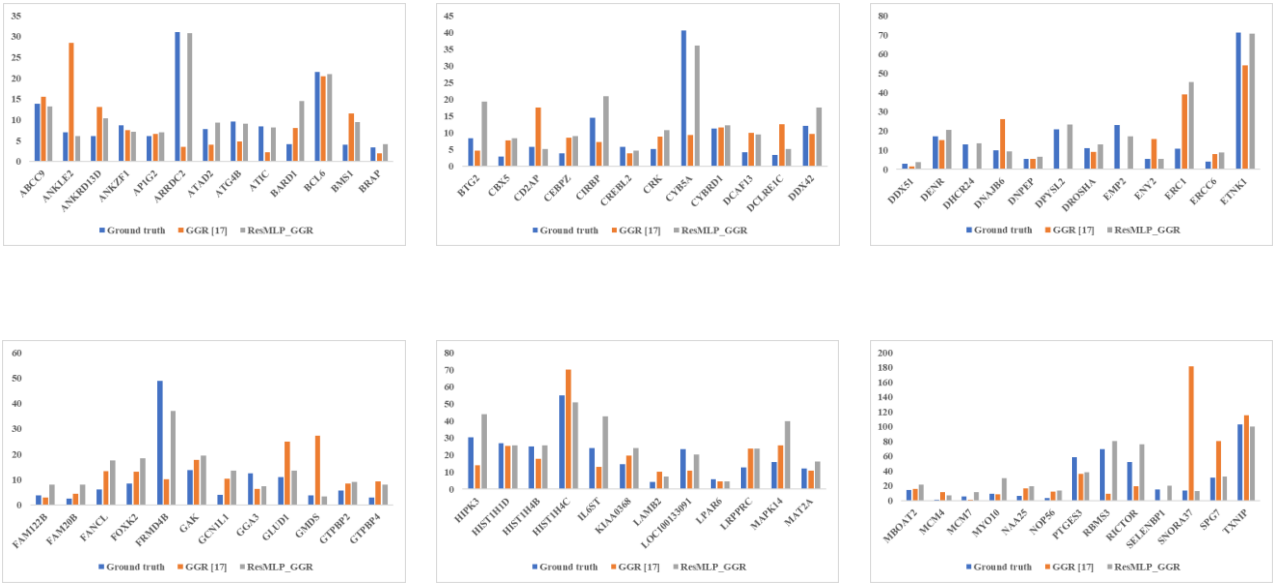


Figure 7. Comparison between estimated gene values (GGR [17], ResMLP_GGR) and actual gene (ground truth) values from one patient as an example. The horizontal axis lists the names of 74 genes of patient, and the vertical axis indicates the FPKM value of each gene.

C. Ablation Study

We experimentally verify the effectiveness of each key component of the proposed method, all using the same control dataset [20]. As shown in the ablation experiments in Table III, for the gene estimation model, the baseline has the MLP structure and estimates the genes using the 74 gene estimation models, and proposed 1 is the same model but uses only one gene estimation model. Different from baseline and proposed 1, ResMLP was substituted for MLP in the baseline to obtain proposed 2. Proposed 3 uses the ResMLP structure and only one gene estimation model. Table III illustrates that when the end-to-end structure (using only one gene estimation model), the recurrence prediction accuracy of the MLP model is improved by 2%. Instead of using an end-to-end structure, when only using the ResMLP module, proposed 2 improved AUC and prediction accuracy (ACC) by 5% and 3%, respectively, over the baseline model, with both sensitivity and specificity above 0.81. Among the four models, proposed 3 achieves the best prediction results, which significantly improves the prediction ability. The ablation experimental results demonstrate that the key components of our proposed method are effective and necessary.

D. Comparison with State-of-the-Art Methods

To verify the performance of our proposed method in NSCLC recurrence prediction, we further compared the proposed method with other methods under the same experimental setup. As shown in Table IV, we applied 10-

fold cross-validation 10 times to measure the average performance of all methods in terms of prediction accuracy (ACC), AUC, sensitivity, specificity, and the number of parameters. The results of the proposed method are bolded. The first four methods in the table are based on handcrafted radiomics features [6–9], the fifth and sixth methods are based on DL [9], and the method in the seventh row is a combination method (combination of deep and handcrafted features) [14]. The eighth and ninth rows are the methods of GGR and GGR_fusion [17, 18], respectively. As shown in Table IV, because the related radiomics or DL method only uses CT images to predict the NSCLC recurrence, its prediction performance is very limited, the prediction accuracy of traditional radiomics is 78.61% (AUC = 0.66), whereas that of the DL-based method (ResNet-50) is 79.09% (AUC = 0.67). In this study, to improve the performance of NSCLC recurrence prediction and meet the demand of all patients with NSCLC, the same as the GGR and GGR_fusion, we are no longer use a single input and single prediction model to directly predict recurrence, but use handcrafted features and deep features, and gene labels first trained in a regression model to fit a gene estimation function. Then, we use a classification prediction model to predict NSCLC recurrence. During testing, based on the mapping relationship between genes and CT images, NSCLC recurrence prediction can be achieved using only CT images.

TABLE III. THE ABLATION STUDY RESULTS OF THE PROPOSED METHOD FOR RECURRENCE PREDICTION

	Estimator's structure	Number of gene estimation models	AUC	ACC	Sensitivity	Specificity
Baseline [17]	MLP for each gene	74	0.77	0.83	0.95	0.59
Proposed 1	MLP for mixed genes	1	0.77	0.85	0.78	0.87
Proposed 2 [19]	ResMLP for each gene	74	0.82	0.86	0.81	0.89
Proposed 3 (ResMLP_GGR)	ResMLP for mixed genes	1	**0.84**	**0.88**	0.85	**0.89**

TABLE IV. PERFORMANCE OF DIFFERENT METHODS TO PREDICT NSCLC RECURRENCE

Methods		ACC (%)	AUC	Sensitivity	Specificity	Params (M)
Wang *et al.* (2019) [6]	PCA+SVM	67.05	0.58	0.85	0.31	-
Lee *et al.* (2020) [7]	Relief-F+SVM	68.18	0.56	0.95	0.14	-
Christie *et al.* (2021) [8]	LASSO	61.36	0.68	0.71	0.62	-
Aonpong *et al.* (2020) [9]	F-test + ANN	78.61	0.66	0.90	0.56	-
Aonpong *et al.* (2020) [9]	ResNet-50	79.09	0.67	0.89	0.59	25.56
Aonpong *et al.* (2020) [9]	DenseNet121	77.36	0.69	0.97	0.38	7.98
Marentakis *et al.* (2021) [14]	Radiomics + DL	82.08	0.71	0.97	0.51	27.16
Aonpong *et al.* (2021) [17]	GGR	83.28	0.77	0.95	0.59	1,679.29
Aonpong *et al.* (2021) [18]	GGR_fusion	84.39	0.79	0.91	0.65	1,701.86
Ai *et. al.* (2022) [19]	ResMLP	86.38	0.82	0.81	0.89	1,146,92
ResMLP_GGR (proposed)	**ResMLP_GGR**	**87.78**	**0.84**	0.85	**0.89**	**17.21**

Different from the GGR and GGR_fusion methods [17, 18], we proposed an end-to-end gene estimation model (which uses only one gene estimation model) based on the ResMLP module. This module can comprehensively integrate multimodal and multi-dimensional data, has strong generalization ability, and can adapt to avoid overfitting while using small datasets. The improved deep MLP recurrence prediction model has strong learning ability, good robustness, high precision and fewer parameters. Unlike the GGR and GGR_fusion methods, a regression model can only be used to estimate one gene. This means that 74 gene estimation models need to be trained for 74 gene estimation, which is very time consuming and inefficient. By utilizing an end-to-end training approach for 74 gene estimation, the proposed method estimates all 74 genes at once using a single regression model (gene estimation model), which significantly reduces the training time. After using the ResMLP-based end-to-end gene estimation model, the number of parameters for the NSCLC recurrence prediction task (including the gene estimation model and the recurrence prediction model) was significantly reduced to 17.21M, which was 8.35M less than that when only using ResNet-50, whereas the number of parameters of the GGR and GGR_fusion method is 1,679.29M and 1,701.86M, respectively. The prediction accuracy of the proposed method was also improved from 83.28% (AUC= 0.77) and 84.39% (AUC= 0.79) of the GGR method and GGR_fusion method to 87.78% (AUC= 0.84). Moreover, it is not necessary to use handcrafted and deep features and estimated gene information simultaneously when performing recurrence prediction training as in GGR_fusion, and only estimated genes can achieve state-of-the-art prediction performance. Fig. 8 depicts the

average Receiver Operating Characteristic (ROC) curves of these different methods, and the results demonstrate the superiority of our proposed method.

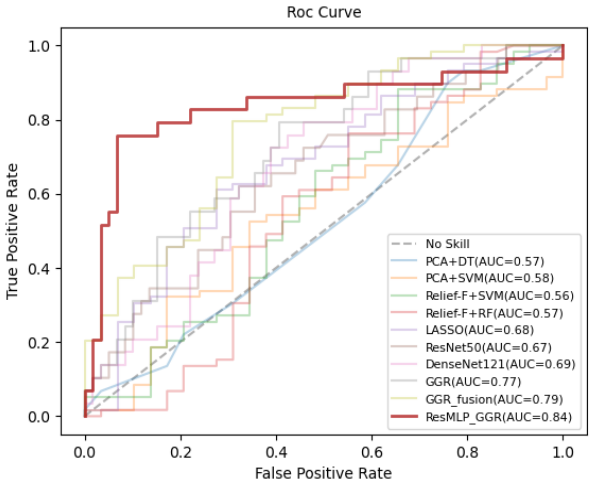


Figure 8. Average ROC curve of NSCLC recurrence prediction uses different methods.

V. Conclusion

This paper proposed a high-precision, residual multilayer perceptrons-based genotype-guided method for NSCLC recurrence prediction, which consists of two models: a gene estimation model and a recurrence prediction model. For the gene estimation model, the ResMLP module is proposed, which uses two residual MLPs as a unit for feature extraction and fusion to build gene estimation functions for multimodal (handcrafted and deep features) inputs. The recurrence prediction model with a deep MLP structure uses the estimation information output from the gene estimation model to predict NSCLC recurrence. According to the experimental results, compared with radiomics methods, DL methods, and the GGR and GGR_fusion methods, the proposed ResMLP_GGR method can more fully mine potential information related to the early recurrence of NSCLC. It exhibits strong fusion and generalization ability, robustness, and high accuracy, and can predict the recurrence probability of all patients with NSCLC. However, note that the proposed method is based on a small public dataset (only 88 patients). Although the regression model we designed has significantly improved the accuracy of NSCLC recurrence prediction, whether these genes (74 genes screened for recurrence prediction) can be used to guide clinical diagnosis still needs to be verified in future studies. In our future work, we plan to apply the proposed algorithm to clinical diagnosis, which will necessitate expanding the training dataset, and focusing on improving the model's prediction performance and generalization ability of NSCLC, bringing its prediction accuracy closer to that of genomics-based methods but using only CT images.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yang Ai provided ideas, programming, conducted the experiments, summarized results and wrote the paper; Yinhao Li provided ideas and revised the paper; Panyanat Aonpong preprocessing; Xianhua Han checked the paper; Yen-Wei Chen checked the paper and provided financial support; all authors had approved the final version.

Funding

This work was supported in part by the Grant in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant No. 20KK0234, No. 21H03470, and No. 20K21821.

References

[1] P. E. Postmus, K. M. Kerr, and M. Oudkerk, "Early and locally advanced Non-Small-Cell Lung Cancer (NSCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 28, no. iv1–iv21, 2017.

[2] H. Uramoto and F. Tanaka, "Recurrence after surgery in patients with NSCLC," *Trans Lung Cancer Res.*, vol. 3, no. 4, p. 242, 2014.

[3] S. R. Deepti, "A survey on application of machine learning algorithms in cancer prediction and prognosis," in *Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing*, N. Sharma, A. Chakrabarti, V. Balas, J. Martinovic, Eds. vol. 1174, Springer, Singapore, 2021, doi: https://doi.org/10.1007/978-981-15-5616-6_25

[4] K. Bera, N. Braman, A. Gupta, *et al.*, "Predicting cancer outcomes with radiomics and artificial intelligence in radiology," *Nat Rev Clin Oncol.*, pp. 1–15, 2021.

[5] T. A. D'Antonoli, A. Farchione, J. Lenkowicz, *et al.*, "CT radiomics signature of tumor and peritumoral lung parenchyma to predict non-small cell lung cancer postsurgical recurrence risk," *Acad Radiol.*, vol. 27, no. 4, pp. 497–507, 2020.

[6] X. Wang, H. Duan, and S. Nie, "Prognostic recurrence analysis method for non-small cell lung cancer based on CT imaging," in *Proc. ICPAI*, Shanghai, China, 2019, 11321T.

[7] S. Lee, J. Jung, H. Hong, *et al.*, "Radiomic feature-based prediction model of lung cancer recurrence in NSCLC patients," in *Proc. IWAIT*, Yogyakarta, Indonesia, 2020, 115150N.

[8] J. R. Christie, M. Abdelrazek, P. Lang, *et al.*, "A multi-modality radiomics-based model for predicting recurrence in non-small cell lung cancer," in *Proc. Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 2021, 116000L.

[9] P. Aonpong, Y. Iwamoto, W. B. Wang, *et al.*, "Hand-crafted and deep learning-based radiomics models for recurrence prediction of non-small cells lung cancers," in *Innovation in Medicine and Healthcare. Smart Innovation, Systems and Technologies*, Y. W. Chen, S. Tanaka, R. Howlett, L. Jain, Eds. vol 192, Springer, Singapore, 2020, doi: https://doi.org/10.1007/978-981-15-5852-8_13

[10] Y. Zhong, Y. She, J. Deng, *et al.*, "Deep learning for prediction of N2 metastasis and survival for clinical stage I non-small cell lung cancer," *Radiology*, vol. 302, no. 1, pp. 200–211, 2022.

[11] Z. Jiao, H. Li, Y. Xiao, *et al.*, "Integration of deep learning radiomics and counts of circulating tumor cells improves prediction of outcomes of early-stage NSCLC patients treated with stereotactic body radiation therapy," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 112, no. 4, pp. 1045–1054, 2022.

[12] J. Yang, J. Ju, L. Guo, *et al.*, "Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 333–342, 2022.

[13] V. Subramanian, N. D. Minh, and S. Tanveer, "Multimodal fusion of imaging and genomics for lung cancer recurrence prediction," in *Proc. ISBI*, Iowa, USA, 2020.
[14] P. Marentakis, P. Karaiskos, V. Kouloulias, *et al.*, "Lung cancer histology classification from CT images based on radiomics and deep learning models," *Med Biol Eng Comput.*, vol. 59, no. 1, pp. 215–226, 2021.
[15] N. Braman, J. W. H. Gordon, E. T. Goossens, *et al.*, "Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data," in *Proc. MICCAI*, Strasbourg, France, 2021, pp. 667–677.
[16] L. Schneider, S. Laiouar-Pedari, S. Kuntz, *et al.*, "Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review," *Eur. J. Cancer*, vol. 160, pp. 80–91, 2022.
[17] P. Aonpong, Y. Iwamoto, X. H. Han, *et al.*, "Genotype-guided radiomics signatures for recurrence prediction of non-small-cell lung cancer," *IEEE Access*, vol. 14420, pp. 90244–90254, 2021.
[18] P. Aonpong, Y. Iwamoto, X. H. Han, *et al*., "Improved genotype-guided deep radiomics signatures for recurrence prediction of non-small cell lung cancer," in *Proc. EMBC*, Mexico, 2021, pp. 3561–3564.
[19] Y. Ai, P. Aonpong, W. Wang, *et al.*, "Residual multilayer perceptrons for genotype-guided recurrence prediction of non-small cell lung cancer," in *Proc. EMBC*, Glasgow, 2022, pp. 447–450.
[20] K. Clark, B. Vendt, K. Smith, *et al.*, "The Cancer Imaging Archive (TCIA), maintaining and operating a public information repository," *J Digit Imaging*, vol. 26, pp. 1045–1057, 2013.
[21] S. Bakr, O. Gevaert, S. Echegaray, *et al*., "A radio genomic dataset of non-small cell lung cancer," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.
[22] V. Van, C. Marcel, *et al.* "RNA-Seq: Revelation of the messengers," *Trends Plant Sci.*, vol. 18, no. 4, pp. 175–179, 2013.
[23] R. Sato, Y. Iwamoto, K. Cho, *et al.*, "Accurate BAPL score classification of brain PET images based on convolutional neural Networks with a joint discriminative loss function," *Appl. Sci.*, vol. 10, 2020
[24] S. Moreno, M. Bonfante, E. Zurek, *et al*., "Study of medical image processing techniques applied to lung cancer," in *Proc. CISTI*, Coimbra, Portugal, 2019, pp. 1–6.
[25] P. Ghosh, A. Karim, S. T. Atik, *et al*., "Expert cancer model using supervised algorithms with a LASSO selection approach," *IJECE*, vol. 11, no. 3, p. 2631, 2021.
[26] M. Wessolly, S. Stephan-Falkenau, A. Streubel, *et al.*, "Digital gene expression analysis of NSCLC-patients reveals strong immune pressure, resulting in an immune escape under immunotherapy," *BMC Cancer*, vol. 22, no. 1, pp. 1–13, 2022.
[27] P. Aonpong, Q. Chen, Y. Iwamoto, *et al*., "Comparison of machine learning-based radiomics models for early recurrence prediction of hepatocellular carcinoma," *J Image Graph.*, vol. 7, no. 4, pp. 117–125, 2019.
[28] P. P. Ohanian, R. C. Dubes, *et al*., "Performance evaluation for four classes of textural features," *CVPR*, vol. 25, no. 8, pp. 819–833, 1992.
[29] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 6, pp. 610–621, 1973.
[30] F. T. Ulaby, F. Kouyate, B. Brisco, *et al*., "Textural infornation in SAR images," *IEEE Trans Geosci Remote Sens.*, vol. 2, pp. 235–245, 1986.
[31] K. He, X. Zhang, S. Ren, *et al*., "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016.
[32] O. Russakovsky, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252. 2015.
[33] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," arXiv preprint, arXiv:1701.09175, 2017.