Research on the Auxiliary Classification and Diagnosis of Lung Cancer Subtypes Based on Histopathological Images

MIN LI[1,2,3], **XIAOJIAN MA**[4], **CHEN CHEN**[5], **YUSHUAI YUAN**[5], **SHUAILEI ZHANG**[5], **ZIWEI YAN**[5], **CHENG CHEN**[5], **FANGFANG CHEN**[5], **YUJIE BAI**[1], **PANYUN ZHOU**[1], **XIAOYI LV**[1,2,3], **AND MINGRUI MA**[4]

[1]College of Software, Xinjiang University, Urumqi 830046, China
[2]Key Laboratory of Software Engineering Technology, Xinjiang University, Urumqi 830046, China
[3]Key Laboratory of Signal Detection and Processing, Xinjiang University, Urumqi 830046, China
[4]The Affiliated Cancer Hospital of Xinjiang Medical University, Urumqi 830011, China
[5]College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

Corresponding authors: Xiaoyi Lv (xjuwawj01@163.com) and Mingrui Ma (119345263@qq.com)

ABSTRACT Lung cancer (LC) is one of the most serious cancers threatening human health. Histopathological examination is the gold standard for qualitative and clinical staging of lung tumors. However, the process for doctors to examine thousands of histopathological images is very cumbersome, especially for doctors with less experience. Therefore, objective pathological diagnosis results can effectively help doctors choose the most appropriate treatment mode, thereby improving the survival rate of patients. For the current problem of incomplete experimental subjects in the computer-aided diagnosis of lung cancer subtypes, this study included relatively rare lung adenosquamous carcinoma (ASC) samples for the first time, and proposed a computer-aided diagnosis method based on histopathological images of ASC, lung squamous cell carcinoma (LUSC) and small cell lung carcinoma (SCLC). Firstly, the multidimensional features of 121 LC histopathological images were extracted, and then the relevant features (Relief) algorithm was used for feature selection. The support vector machines (SVMs) classifier was used to classify LC subtypes, and the receiver operating characteristic (ROC) curve and area under the curve (AUC) were used to make it more intuitive evaluate the generalization ability of the classifier. Finally, through a horizontal comparison with a variety of mainstream classification models, experiments show that the classification effect achieved by the Relief-SVM model is the best. The LUSC-ASC classification accuracy was 73.91%, the LUSC-SCLC classification accuracy was 83.91% and the ASC-SCLC classification accuracy was 73.67%. Our experimental results verify the potential of the auxiliary diagnosis model constructed by machine learning (ML) in the diagnosis of LC.

INDEX TERMS Lung cancer, histopathological images, computer-aided diagnosis, lung adenosquamous carcinoma (ASC), machine learning.

I. INTRODUCTION

Lung cancer (LC) is one of the most common malignant tumors worldwide [1]–[3]. According to the 2018 International Agency for Research on Cancer statistics, there will be 2.1 million new cases of LC and 1.8 million deaths worldwide [4]. Due to its high morbidity and mortality, it has become one of the most serious cancers threatening human health. Clinicians' visual analysis of LC histopathological

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

images is one of the most important methods for evaluating LC subtypes [5]. However, it is complicated and challenging for pathologists to review thousands of histopathological images, and it is even more difficult for doctors with less experience [6]. Therefore, to relieve the pressure on doctors and improve the accuracy and efficiency of diagnosis, it is particularly important to study the computer-aided diagnosis model of LC [7].

From the perspective of pathology and treatment, LC can be divided into non-small cell lung carcinoma (NSCLC) and small cell lung carcinoma (SCLC), of which 80%-85% are

NSCLC and the rest are SCLC [8]. The main histological types of NSCLC are lung adenocarcinoma (ADC) and lung squamous cell carcinoma (LUSC). The other histological types of NSCLC are lung adenosquamous carcinoma (ASC), large-cell carcinoma [9]. In particular, ASC is a relatively rare subtype of NSCLC that accounts for 0.3–5% of all NSCLCs [10]. Due to the different histopathological types of LC, the treatment methods adopted are also different. When the lung tissue classification is determined, the appropriate treatment mode can be selected, such as the reasonable application of surgery, chemotherapy, radiotherapy, molecular targeted therapy and immune therapy. In addition, LC screening errors can be avoided, clinicians' multifarious work pressure can be slowed, patients' survival time can be maximized and the patient's quality of life improved.

The LC imaging examination methods mainly include: (1) X-ray photography, which is one of the most basic lung imaging examination methods, but the resolution of the photography is low and there are blind spots in the examination [11]. (2) Computed tomography (CT), specifically, chest CT has advantages in detecting early peripheral LC and identifying the location of the lesion. Currently, it is one of the most commonly used imaging methods for preoperative diagnosis and staging of LC [12]. However, one of the limitations of using CT examinations is that for patients undergoing repeated examinations, it is necessary to consider the impact of the radiation dose produced by the operation [13]. (3) Magnetic resonance imaging (MRI) has high sensitivity and specificity for vertebral and bone metastases, but it is not recommended for routine diagnosis of LC [13], [14]. (4) Ultrasound is a non-invasive tool, which is usually superior to radiography in the examination of postoperative pulmonary complications (PPCs). It has developed into a valuable method [15]. Although the above imaging examination methods play an important role in detecting of LC, the results of each examination are only used as a reference for the diagnosis, staging, re-staging, efficacy monitoring and prognosis evaluation of LC, while histopathological examination is the gold standard for tumor qualitative and clinical staging [16], [17].

However, due to the complex texture features of histopathological images, as far as the authors know, there is no computer-aided diagnosis method for ASC, LUSC and SCLC based on histopathological images. To fill this gap, this paper includes ASC sample data for the first time and introduces a computer-aided model for automatic classification of LC subtypes based on histopathological images of LC. First, seven texture analysis methods are used to extract 265-dimensional features of LC histopathological images, and the relevant features (Relief) algorithm is used for feature selection. Compared with a variety of machine learning (ML) models, the Relief-SVM model obtains optimal performance for LC subtype classification [18]. The main contributions of this study are as follows:

- This study was the first to include relatively rare ASCs in lung histopathological images and applies them to the automatic classification of LC subtypes.
- This paper was the first to apply the Relief-SVM algorithm to the classification of LC histopathological images, which demonstrates the tremendous potential of ML algorithms to be used in the diagnosis of LC subtypes.
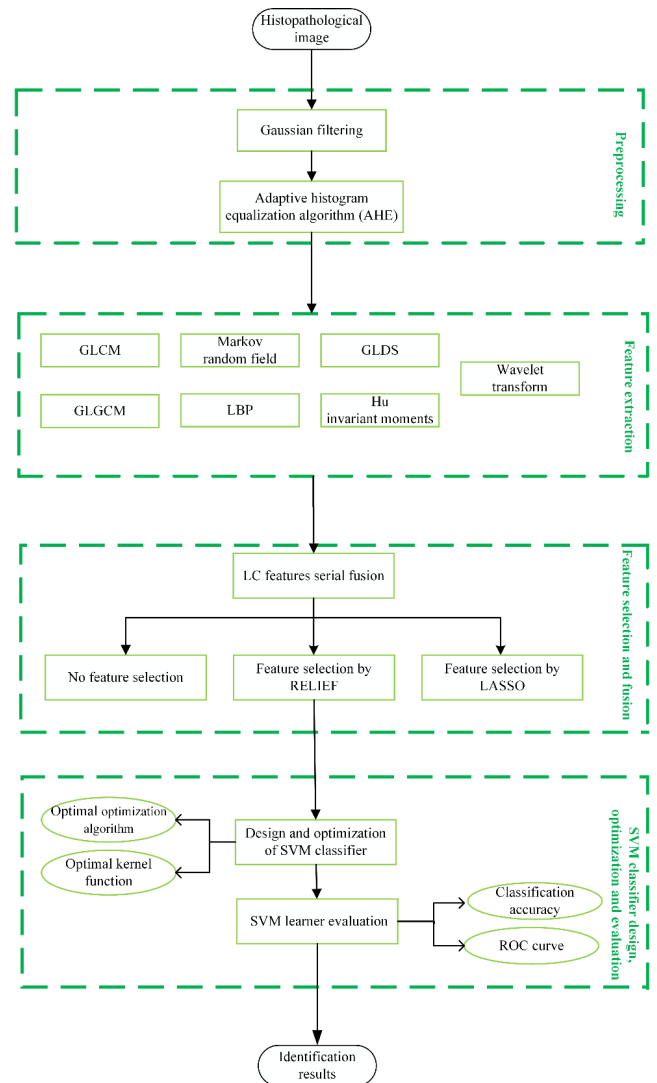

FIGURE 1. Overall flowchart of the auxiliary classification and diagnosis of LC subtypes. Explain some abbreviations in the figure: gray-level gradient co-occurrence matrix (GLGCM), gray-level co-occurrence matrix (GLCM), local binary pattern (LBP), gray-level difference statistics (GLDS), the least absolute shrinkage and selection operator (LASSO).

Figure 1 shows the overall flow chart of the auxiliary classification and diagnosis of LC subtypes. In the following sections, we will introduce them in detail.

II. MATERIALS AND METHODS
A. DATA COLLECTION
This study included 94 patients from a Cancer Hospital in Xinjiang from October 2011 to October 2020, of whom
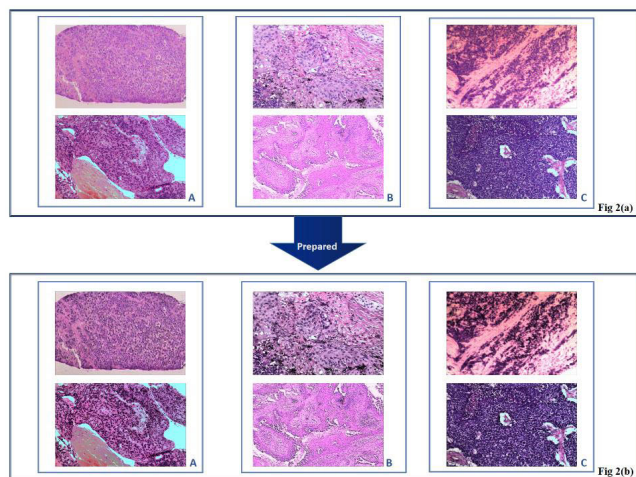
FIGURE 2. 2(a). The unprocessed histological images of LUSC, ASC and SCLC. 2(b). The processed histological images of LUSC, ASC and SCLC. To better display the image details, two pathological tissue images of each category were selected for display in this study, of which A was LUSC, B was ASC, and C was SCLC.

22 had LUSC, 27 had ASC and 45 had SCLC. All subjects had signed informed consent in this retrospective study. For improving the generalization performance of the model, this study took an average of 1-2 pathological tissue sections from each patient as the research object, and finally selected 121 histopathological images [19], [20]. According to the classification standards provided by the World Health Organization, LC patients were collected from the hospital pathology database, and the relevant pathological diagnosis results were confirmed by pathology. The detailed information is shown in Table 1. All tumor tissues included in the study were made into histopathological sections by hematoxylin and eosin (H&E) staining [21], which were collected through the microscope of the hospital pathology department and stored as JPG image files with four resolutions: 744×554, 2048×1536, 1024×768, 640×480 [6].

TABLE 1. Patient emographics (n = 94).

Characteristic LC	Sex		Age	
	Men	Women	Mean	Range
LUSC	21	1	64.95	[48,85]
ASC	15	12	62.19	[45,76]
SCLC	33	12	60.78	[40,81]

B. HISTOPATHOLOGICAL IMAGE PREPROCESSING

Histopathological examination is the gold standard for qualitative and clinical tumor staging [22]. Histopathological images have been widely used by doctors for diagnosis and treatment, and are an important basis for predicting patient survival [23]. The histopathological images of LUSC, ASC

and SCLC are shown in Figure 2. According to reports, the following problems exist in histopathological images:

- The histopathological images are faced with a large number of rich geometric structures and complex textures caused by the diversity of structural morphology [24].
- The histopathological images are easily affected by color differentiation and noise due to external reasons such as illumination conditions [25].
- Due to the difference in microscope magnification, equipment parameters and other factors in histopathological images, the image size and resolution are not the same [6].
- Texture features such as local microvessels in histopathological images are the key to disease diagnosis, and the extraction of features is of great significance to assist in the diagnosis and classification of LC [26].

For all these reasons, the histopathological images presented to us are often not perfect. As an efficient low-pass filter, the Gaussian filter has shown good filtering performance in both the spatial domain and frequency domain, and is widely used in image processing to reduce noise [27]. Generally, the pixel value of each point is replaced by the weighted mean of its neighbourhood, so Gaussian filtering is the process of weighted averaging of the whole image. The histopathological images processed by Gaussian filtering tend to be smoother and contain less noise, which is ready for our subsequent studies [28]. According to the Ciompi *et al.* description, the existence of differences such as color destandardization in histopathological images will limit the interpretation of histopathological images by inexperienced pathologists, and the color differences will affect the performance of the automatic diagnosis model [25]. To avoid the problem of information loss due to excessive brightness of histopathological images [26], we adopt the adaptive histogram equalization algorithm (AHE) to normalize the color of histopathological images in this study [26]. In this mode, the image is divided into several 8×8 pieces, and then the histograms of multiple local regions of the image segmentation are calculated. At this point, the brightness is redistributed in the 8×8 area to change the image contrast. Therefore, this algorithm is more beneficial for improving the local contrast of the image and obtain more image details [26]. According to the investigation, the improvement in model classification performance by the color normalization method is limited. The image after the above image preprocessing operation is shown in Figure 2(b). Figure 3 shows the details before and after the preprocessing operation of the ASC histopathological image.

C. FEATURE EXTRACTION OF HISTOPATHOLOGICAL IMAGES OF LUNG CANCER

Because the rich features presented in histopathological images are an important basis for clinicians to carry out diagnosis, the effective extraction of image features is the key to improving the accuracy of computer-aided diagnosis [29].
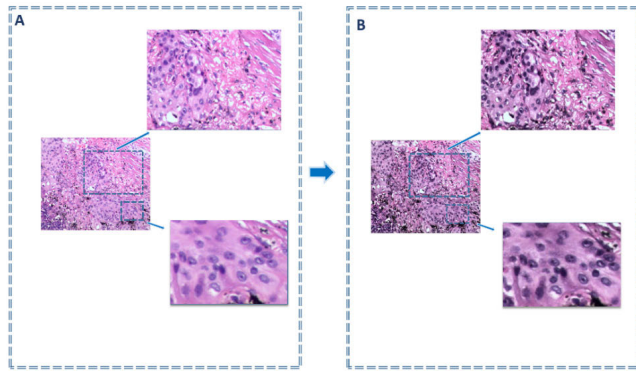
FIGURE 3. Details before and after the preprocessing operation of the ASC histopathological image. A. Details before the preprocessing operation of the ASC histopathological image. B. Details after the preprocessing operation of the ASC histopathological image.

FIGURE 4. LBP gray-coded restored image. (a) LUSC, (b) ASC, (c) SCLC.

This paper explores the effect of two feature extraction methods on LC histopathological image classification.

1) HANDCRAFTED TEXTURE EXTRACTION METHODS

We extracted 265-dimensional features using seven handcrafted texture extraction algorithms, including the Hu invariant moments, GLGCM, wavelet transform, GLCM, LBP, GLDS and Markov random field, as shown in Table 2 [18].

TABLE 2. Extracted features of the seven methods.

Methods	Total number	Methods	Total number
Hu invariant moments	7	GLDS	4
Markov random field	12	GLCM	8
GLGCM	15	Wavelet transform	160
LBP	59		

a: LBP

The LBP method for texture feature extraction was proposed in 1994 [30]. However, the greatest defect of the traditional LBP algorithm cannot satisfy the requirement of textures of different sizes and frequencies. It only covers a small area within a fixed radius, which makes it impossible to extract the features of the whole histopathological image perfectly. To compensate make up for this deficiency, and to meet the requirements of a constant grey level and rotation, this paper uses a circular neighbourhood to replace the traditional square neighbourhood LBP algorithm [31]. The LBP algorithm has the advantages of simple calculation, efficient recognition, good texture feature display and low computational complexity [32], [33].

As the traditional LBP algorithm cannot efficiently extract the texture features from the large-scale image structure, the improved LBP operator in the circular domain is adopted in this study to allow the existence of N pixels in the circular domain with a radius of R. Using the above theory, LBP feature processing was performed on the histopathological
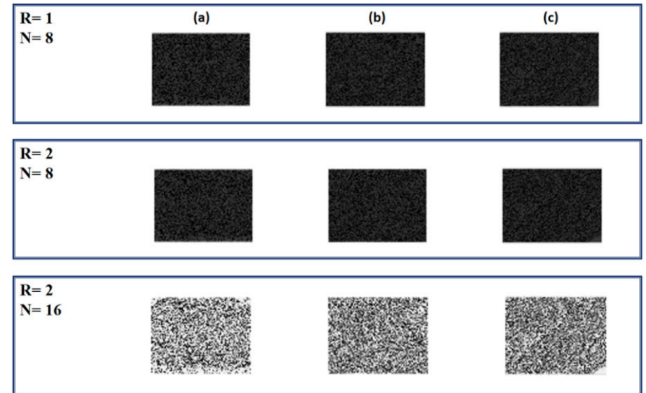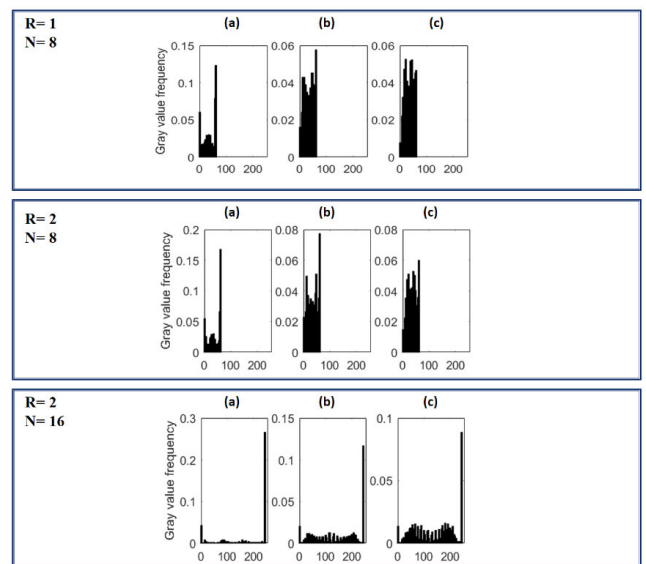

FIGURE 5. Histogram of gray statistics of histopathological images. (a) LUSC, (b) ASC, (c) SCLC.

images, and the results are shown in Figure 4 and Figure 5. As seen from Figure 4 and Figure 5, when R = 1 and N = 8 schemes and R = 2 and N = 8 schemes, the grey values of the LBP algorithm are mostly concentrated in the 0-100 threshold range. The grey values distribution of the R = 2 and N = 16 schemes was relatively uniform. However, the experiments showed that the extracted 59-dimensional features were more helpful in distinguishing histopathological images of LC when R = 2 and N = 8, while the extracted features had more redundant information when R = 2 and N = 16, which was not conducive to the improvement of the subsequent classification effect.

b: GLCM

GLCM is a regular method for describing the texture features by studying the spatial correlation characteristics of grey images. Calculating the joint probability density of two pixels positions to generate the co-occurrence matrix, the co-occurrence matrix is statistical second-order statistical
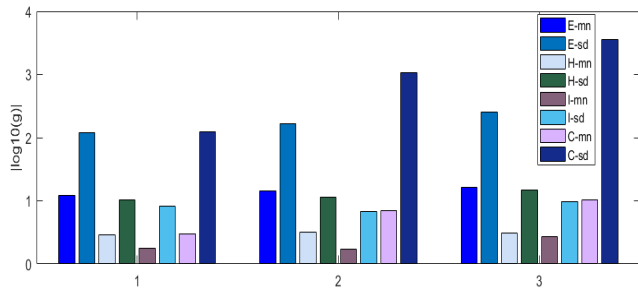
FIGURE 6. Histogram of eigenvalues of the GLCM. 1. LUSC, 2. ASC, 3. SCLC.

FIGURE 7. Histogram of eigenvalues of the GLGCM. 1. LUSC, 2. ASC, 3. SCLC.

characteristics of image brightness changes; it not only reflects the characteristics of the distribution of brightness, it also reflects the same or close to the brightness of the pixel brightness between the location of the distribution characteristics, thus constituting a set of texture features to better show the image features [34], [35]. However, due to the existence of a large number of co-occurrence matrices, it is generally not used directly in the field of image processing. Instead, texture feature quantities are extracted again according to the grey co-occurrence matrices. The commonly used local steady characteristic is the energy E, entropy H, moment of inertia of I, and correlation C. Thus, this study will consider these four with the mean and standard deviation for subsequent image recognition and the classification statistic formula (such as (1)-(4)), and three points in the lung cancer model experiment on the image to extract the characteristics of the above four mean (MN) and standard deviation (SD) as shown in Figure 6.

$$\text{Energy}: \quad E\,(d,\theta)$$
$$= \sum_{i,j} \{p\,(i,j\,|\,d,\theta)\}^2 \tag{1}$$
$$\text{Entropy}: \quad H\,(d,\theta)$$
$$= -\sum_{i,j} \{p\,(i,j\,|\,d,\theta)\} - log\,\{p\,(i,j\,|\,d,\theta)\} \tag{2}$$
$$\text{Moment of inertia}: \quad I\,(d,\theta)$$
$$= \sum_{i,j} \{i,j\}^2\, p\,(i,j\,|\,d,\theta) \tag{3}$$
$$\text{Correlation}: \quad C\,(d,\theta)$$
$$= \frac{\sum_{i,j} (i-\mu_i)\,(j-\mu_j)\,p\,(i,j\,|\,d,\theta)}{\sigma_x \sigma_y} \tag{4}$$

c: GLGCM

The grey-gradient co-occurrence matrix synthesizes the grey level and gradient information existing in the image; that is, the image gradient information is added to the grey-level co-occurrence matrix so that the gradient information is mixed in the grey-level co-occurrence matrix and the image feature extraction effect is often better [31], [36]. Based on the standardization of the grey-gradient co-occurrence matrix, we can calculate a series of secondary statistical characteristics. Fifteen is adopted in this study for the commonly used digital features: small grads dominance, big grads dominance, gray asymmetry, grads asymmetry, energy, gray mean, grads mean, gray variance, grads variance, correlation, gray
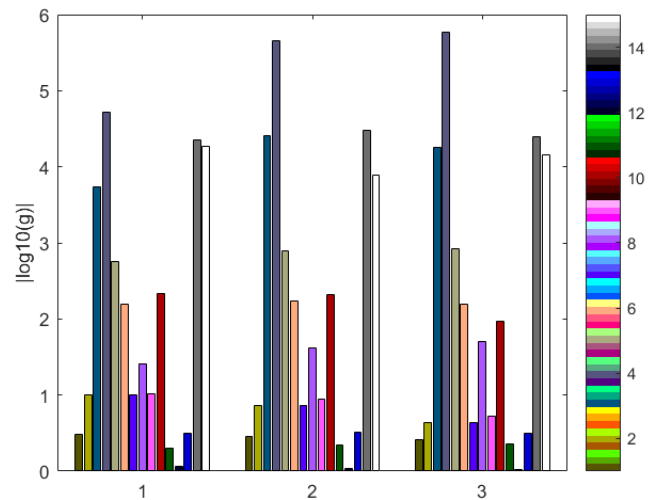
entropy, grads entropy, entropy, inertia, differ moment. After processing the GLGCM on lung cancer histopathological images, fifteen features were extracted from each type of sample and are displayed in Figure 7.

d: OTHER HANDCRAFTED TEXTURE EXTRACTION METHODS

The moment feature mainly expresses the geometric features in the image area. It has invariant characteristics such as rotation, translation, and scale, and is often called an invariant moment [37], [38]. The moment invariant function has been widely used in image pattern recognition, classification, target recognition and other tasks. Therefore, this paper uses Hu invariant moments for the feature representation of lung cancer histopathological images. The seven features of torque change describe the shape features in the tissue images [38]. The gray difference statistical method is one of the detection methods based on statistical texture features [39]. It describes the gray level changes between each pixel of the texture image and its neighboring pixels. The method is simple and easy to implement. In addition, texture feature extraction methods such as Markov random field and wavelet transform are becoming more and more active in the field of pattern recognition [40]–[42].

2) CNN AUTOMATICALLY EXTRACTS HISTOPATHOLOGICAL IMAGE FEATURES OF LUNG CANCER

CNNs have gradually developed into one of the mainstream algorithms in the field of computer vision after decades of development [43], [44]. The back-propagation strategy proposed by LeCun *et al.* in 1990 has always been regarded as the origin of CNN [45]. With the AlexNet network structure proposed by Krizhevsky *et al.*, the AlexNet structure that focuses more on details and complexity and has better performance stands out [46]. It won the 2012 ImageNet competition champion and made a major breakthrough in the image recognition task, which had a strong impact on the field of
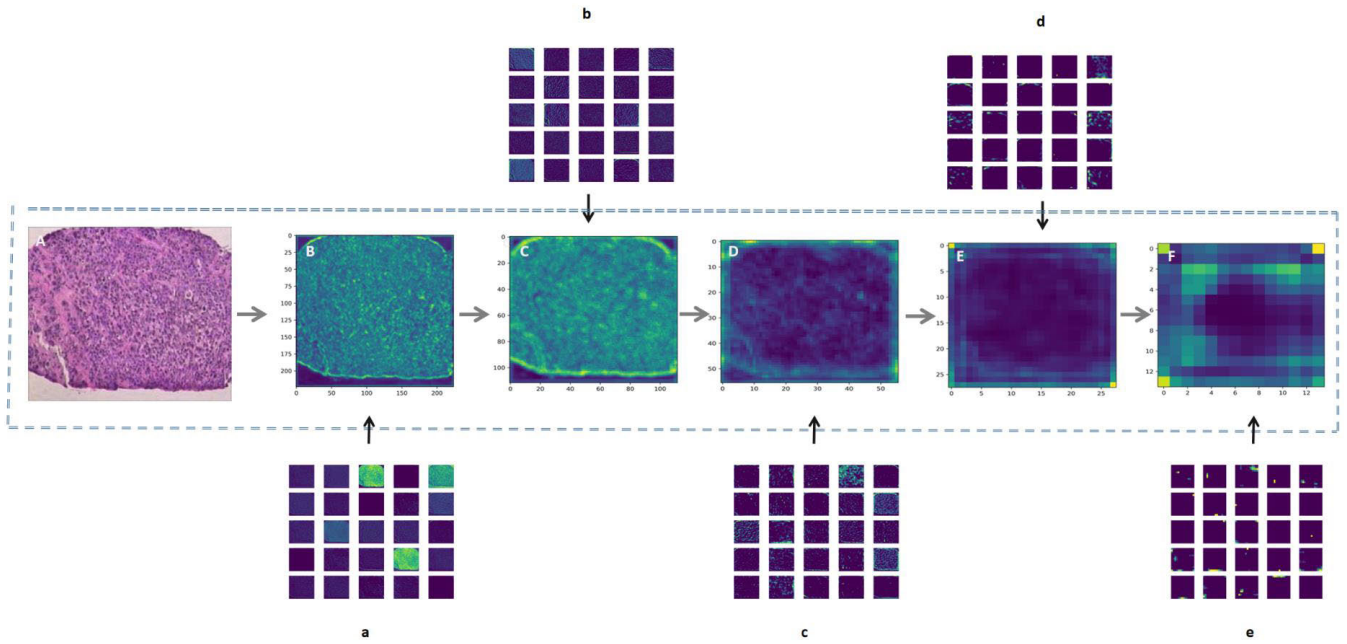
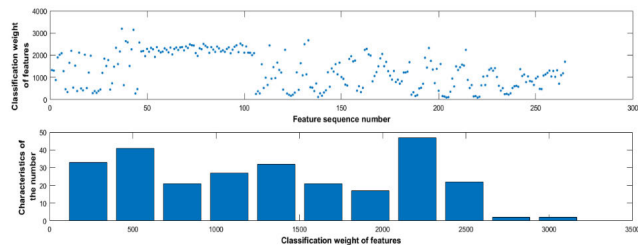**FIGURE 8.** VGG16 model layer by layer feature extraction.

FIGURE 9. Feature classification weight diagram before threshold setting (LUSC-ASC).

computer vision at that time. With the gradual improvement of people's matching structure, CNN has been applied to different fields of human life [47]–[50], with remarkable effects in image classification, target detection, face recognition, natural language processing and other related fields [51]. To date, the modern CNN architecture mainly consists of the following five parts: the convolution layer, the pooling layer, the activation function, the dropout rate (optional) and the full connection layer.

CNN's feature extraction process and the doctor's diagnosis have a certain similarity, and those closer to the bottom of the layers in the convolution-based coding tend to learn more general features and the extracted features can be partial, with high generality characteristic figures and edges, such as visual features (such as color and texture), and those closer to the top of the layer are more abstract with more specialized characteristics. In Figure 8, this study shows the layered features extracted from lung cancer pathological tissue images using VGG16 pairs to verify the reusability of the underlying features [6].
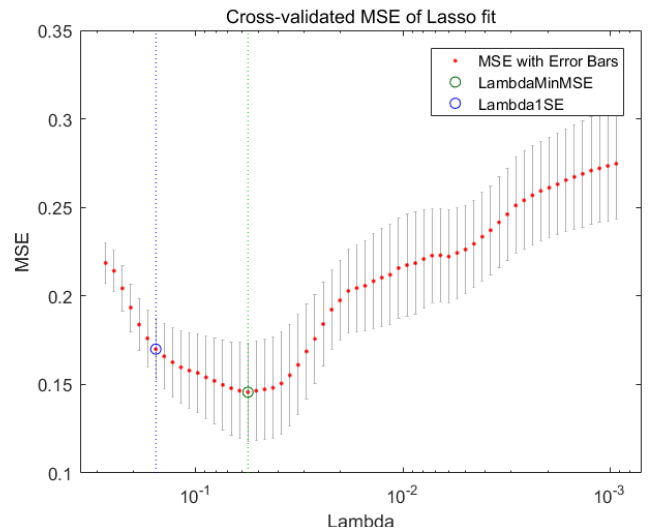

FIGURE 10. Cross-validated MSE of LASSO fit (LUSC-ASC).

D. FEATURE SELECTION, FEATURE DIMENSION REDUCTION AND FUSION

The 265-dimensional features were fused together by serial fusion. Problems such as the dimensionality disaster caused by too many sample attributes will adversely affect the performance of the model [52]. Therefore, it is extremely important to retain key features and eliminate irrelevant features. Feature selection often reduces the difficulty of learning tasks and is more conducive to improving operating efficiency [53]. Relief is a well-known filtering feature selection method. This method designs a "correlation statistic" to measure the importance of features. The LASSO

TABLE 3. Comparing no feature selection, three optimization algorithms (OAs) and four kernel functions (KFs) to choose the optimal combination.

(Test/ training(%)/time) KF / OA	Linear kernel	Polynomial kernel	Gaussian kernel	Sigmoid kernel
LUSC-ASC GS-SVM	73.91/100.00/2.81	73.91/100.00/2.70	73.91/92.16/2.72	65.22/68.63/2.71
PSO-SVM	70.43/98.43/1.22	73.91/100.00/1.37	69.13/83.73/1.41	65.22/68.63/1.45
GA-SVM	71.30/98.43/15.35	73.91/100.00/15.91	66.52/84.31/15.42	65.22/68.63/15.56
LUSC-SCLC GS-SVM	56.52/97.92/2.64	60.87/97.92/2.49	78.26/93.75/2.51	65.22/66.67/2.45
PSO-SVM	58.26/96.25/1.22	59.57/93.96/1.20	74.35/96.25/1.21	65.22/66.67/1.21
GA-SVM	61.30/83.33/14.13	59.57/93.96/13.35	65.22/69.79/13.53	65.22/66.67/13.56
ASC-SCLC GS-SVM	70.00/92.54/4.53	76.67/100.00/4.57	73.33/77.61/4.67	50.00/52.24/4.48
PSO-SVM	68.33/97.91/2.02	75.33/97.61/2.03	73.33/76.12/2.08	50.00/52.24/2.05
GA-SVM	73.67/97.31/25.33	76.33/97.61/26.32	57.67/97.76/25.99	50.00/52.24/24.86

TABLE 4. Comparing LASSO feature selection, three optimization algorithms (OAs) and four kernel functions (KFs) to choose the optimal combination.

(Test/ training(%)/time) KF / OA	Linear kernel	Polynomial kernel	Gaussian kernel	Sigmoid kernel
LUSC-ASC GS-SVM	59.13/86.67/0.38	64.35/90.59/0.34	65.22/89.02/0.42	65.22/68.63/0.36
PSO-SVM	59.13/86.08/0.24	62.17/78.24/0.30	56.52/86.27/0.30	62.17/78.82/0.34
GA-SVM	66.09/90.98/1.56	57.39/94.90/1.53	51.30/93.73/1.47	65.22/68.82/1.46
LUSC-SCLC GS-SVM	65.22/66.67/0.30	65.22/66.67/0.25	65.22/66.67/0.25	65.22/66.67/0.29
PSO-SVM	65.22/66.67/0.29	65.22/66.67/0.25	65.22/66.67/0.24	65.22/66.67/0.25
GA-SVM	65.22/66.67/1.57	65.22/66.67/1.64	65.22/66.67/1.77	65.22/66.67/1.55
ASC-SCLC GS-SVM	50.00/82.99/0.47	50.00/82.54/0.50	50.00/83.28/0.42	46.33/72.54/0.43
PSO-SVM	52.00/82.99/0.40	50.00/55.07/0.35	50.00/83.13/0.37	50.00/81.34/0.36
GA-SVM	50.00/83.43/1.69	50.00/85.52/1.74	50.00/84.63/1.97	51.33/43.43/1.81

algorithm is more conducive to reducing the risk of over-fitting and obtaining a "sparse" solution [54]–[60]. Therefore, this study compared the no feature selection, LASSO feature selection, the relief feature selection model performance and the linear discriminant analysis (LDA) feature dimension reduction [61], [62]. Through a contrast experiment, we chose the relief feature selection. The relief feature selection diagram was shown in Figure 9. Figure 10 shows the LASSO feature selection algorithm using 10-fold cross validation. The Lasso algorithm formula was shown in equation 5.

$$min_{\mathbf{w}} \sum_{i=1}^{m} \left(y_i - w^T x_i \right)^2 + \lambda \| \mathbf{w} \|_1 \qquad (5)$$

E. CLASSIFIER DESIGN AND IMPLEMENTATION

Support vector machines (SVMs) have gradually developed into the mainstream technology of machine learning and have been widely used in every field of life [63]. The basic SVM model is often defined as finding the hyperplane with the largest interval that is most suitable for the classification of samples in the feature space, which makes the constructed linear classifier have stronger generalization performance. Based on the investigation, it was found that the kernel function can better map low-dimensional feature data to high-dimensional data, and the kernel function applied in different fields is not the same. To make features more easily separated or better structured, we carried out a detailed experimental comparison of four kernel functions and the three optimization algorithms in the following experiments to make the SVM learner achieve the optimal classification effect [64], [65].

In addition, we also conducted parallel comparisons with the Back Propagation (BP) Neural Network Classifiers and K-Nearest Neighbor (KNN), Decision tree [29],

TABLE 5. Comparing relief feature selection, three optimization algorithms (OAs) and four kernel functions (KFs) to choose the optimal combination.

(Test/ training(%)/time) KF / OA		Linear kernel	Polynomial kernel	Gaussian kernel	Sigmoid kernel
LUSC-ASC	GS-SVM	73.91/96.08/1.83	73.91/100.00/1.85	73.91/90.20/1.83	65.22/69.63/1.82
	PSO-SVM	71.30/95.29/0.85	73.91/100.00/0.84	71.30/88.63/0.86	65.22/68.63/1.04
	GA-SVM	72.17/96.08/5.10	73.91/100.00/4.66	65.65/90.59/10.87	65.22/68.63/7.38
LUSC-SCLC	GS-SVM	82.61/93.75/1.00	86.97/97.92/0.99	78.26/85.42/0.98	65.22/66.67/1.00
	PSO-SVM	82.17/92.29/0.65	83.91/95.21/0.54	75.22/78.13/0.59	65.22/66.67/0.61
	GA-SVM	76.96/82.92/5.94	83.48/93.75/5.82	69.57/81.88/6.18	65.22/66.67/6.04
ASC-SCLC	GS-SVM	66.67/100.00/2.61	73.33/100.00/2.57	76.67/100.00/2.62	50.00/52.24/2.60
	PSO-SVM	67.33/96.57/1.29	73.67/98.51/1.42	75.33/82.39/1.26	36.00/32.39/1.46
	GA-SVM	67.33/96.57/15.39	73.33/100.00/13.81	56.33/76.12/15.50	79.00/51.04/7.16

Naive Bayesian (NB) classifier [13] to find the best classification model.

F. PERFORMANCE EVALUATION

For the purpose of evaluating the reliability of the model, the receiver operating characteristic (ROC) curve was drawn to evaluate the effect of the classification model, and the area under the curve (AUC) value was quantitatively displayed [66]. According to their true category and learners' predicted category, the samples were divided into the following four conditions: true positive (TP), true negative (TN), false negative (FN) and false positive (FP). The vertical axis of the ROC curve represents the true positive rate (TPR) and the horizontal axis of the ROC curve represents the false positive rate (FPR), defined as Equations (6) and (7), respectively.

$$TPR = \frac{TP}{TP + FN} \quad (6)$$

$$FPR = \frac{FP}{TN + FP} \quad (7)$$

III. EXPERIMENTS AND RESULTS

The experiment in this study was based on the PyCharm platform and MATLAB R2016a platform [67], [68]. The LIBSVM toolbox was used for statistical analysis of the SVM algorithm [69]. All deep learning experiments were implemented in the KERAS deep learning framework based on the TensorFlow 2.0.0 model by using the Python3.6 programming language [70], [71]. Handcrafted texture extraction methods and SVM support vector machine classification were deployed in MATLAB R2016a platform. The experiments in Tables 3, 4, 5 and 6 were repeated ten times, and the average value was used as the experimental result.

A. SELECTION OF THE FEATURE SELECTION METHOD, KERNEL FUNCTION AND OPTIMIZATION ALGORITHM

In this study, no feature selection, LASSO feature selection, Relief feature selection and LDA feature dimension reduction were compared, and four kernel functions and three optimization algorithms were used to attempt to construct the best auxiliary diagnostic model. The three optimization algorithms include grid search-SVM (GS-SVM), particle swarm optimization-SVM (PSO-SVM), and genetic algorithms based on SVM (GA-SVM). The experimental results are shown in Tables 3, 4, 5 and 6. The experimental parameters are shown in Table 7 [18]. According to the comparison of experimental results, the optimal combination of the Relief feature selection, particle swarm optimization algorithm and polynomial kernel was finally selected in this study for the subsequent classification and diagnosis of LUSC, ASC and SCLC.

B. COMPARISON OF THE CLASSIFICATION EFFECT OF LC HISTOPATHOLOGICAL IMAGES WITH OTHER MODELS

1) COMPARISON WITH OTHER CLASSIFIERS

In this study, several mainstream classification algorithms, such as the NB classifier and KNN classifier, were compared with the Relief-SVM model. As seen in Table 8, the Relief-SVM model used in this paper can achieve an overall accuracy of more than 70%, and the classification accuracy of LUSC-SCLC achieved 83.91%. This is a satisfactory result, which can help the application of cytology in the diagnosis of lung cancer. In addition, the classification accuracy achieved by the Relief-SVM model is 27.39% higher than the NB model in LUSC-SCLC and 24.34% higher than the SVM model. We drew the ROC curves of the Decision Tree, BP, KNN, SVM, LASSO-SVM, RELIEF-SVM,

TABLE 6. Comparing LDA feature selection, three optimization algorithms (OAs) and four kernel functions (KFs) to choose the optimal combination.

(Test/ training(%)/time) / KF OA		Linear kernel	Polynomial kernel	Gaussian kernel	Sigmoid kernel
LUSC-ASC	GS-SVM	65.22/68.63/0.35	34.78/72.55/0.37	65.22/82.35/0.38	65.22/68.63/0.40
	PSO-SVM	65.22/72.16/0.31	40.87/71.57/0.34	65.22/86.08/0.40	65.22/68.63/0.33
	GA-SVM	65.22/70.78/1.73	37.83/71.57/1.69	65.22/84.31/1.71	65.22/68.63/2.11
LUSC-SCLC	GS-SVM	65.22/66.67/0.29	65.22/66.67/0.28	65.22/66.67/0.30	65.22/66.67/0.26
	PSO-SVM	65.22/66.67/0.27	62.17/64.38/0.25	65.22/66.67/0.27	65.22/66.67/0.26
	GA-SVM	65.22/66.67/1.72	56.62/61.04/1.70	65.22/66.67/1.82	65.22/66.67/1.74
ASC-SCLC	GS-SVM	50.00/52.24/0.38	50.00/73.13/0.38	50.00/73.13/0.40	50.00/52.24/0.40
	PSO-SVM	50.00/54.33/0.33	50.00/73.13/0.30	50.00/73.13/0.34	50.00/52.24/0.32
	GA-SVM	50.00/66.87/1.96	50.00/62.69/1.99	50.00/73.13/2.12	50.00/52.24/2.07

TABLE 7. The experimental parameters of GS-SVM, PSO-SVM and GA-SVM.

Classifier	Parameter	Parameter value	Parameter	Parameter value
GS-SVM	Range of C	$[2^{-4},\ 2^{4}]$	Range of g	$[2^{-4},\ 2^{4}]$
	Step length of C	0.3	Step length of g	0.3
PSO-SVM	Range of C	$[10^{-1},\ 10^{2}]$	Range of g	$[10^{-2},\ 10^{3}]$
	Local search capability	1.5	Global search capability	1.7
	Maximum number of generations	50	Maximum number of populations	5
GA-SVM	Range of C	$[0,10^{2}]$	Range of g	$[0,10^{3}]$
	Maximum number of generations	200	Maximum number of populations	20
	Generation gap	0.9		

LDA-SVM and NB as shown in Table 9 and the AUC values are shown in Table 10. As seen in Table 10, the AUC value of Relief-SVM for LUSC-SCLC classification increased by 0.3625 compared with BP and 0.3250 compared with LASSO-SVM, and the AUC value of Relief-SVM for LUSC-ASC classification increased by 0.2208 compared with KNN. The results show that Relief-SVM has good feasibility.
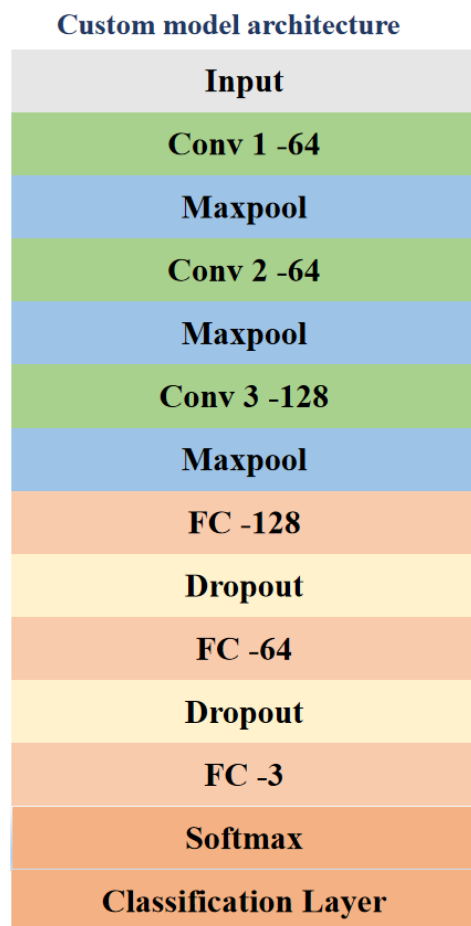
2) COMPARISON WITH CNN

To make our model robust to feature transformation, firstly, the individual data set is randomly divided into 70% for training and 30% for verification. Then, the method of flipping, translation, and multiangle rotation was used to expand the original data for 100 times respectively [72]–[75]. Table 11 shows the number of images in the augmented training set and validation set. Inspired by the researchers [76]–[81], we trained the CNNs model, fine-tuning the pre-trained CNNs model, and custom the CNNs model to conduct experiments on the data set after data augmentation [6], [82].

The custom model architecture were shown in Fig 11. The main experimental parameters of fine-tuning the pre-trained CNNs are shown in Table 12 [6], [78]. Figures 12, 13, 14, and 15 respectively show the accuracy and loss rate curves under different experimental backgrounds. It can be seen from Figures 12 and 13 that the accuracy and loss curves of the four classic CNNs have large fluctuations. Under the guidance of Liu *et al.* [6], although there are large differences between the natural images contained in the ImageNet and medical images, according to the survey, fine-tuning pre-trained CNNs models can reduce the problem of the limited number of samples to a certain extent, and obtain good classification performance. We adjust the model by adding dropout, l2 regularization in the experiment, using different optimization algorithms, and adjusting the learning rate, we show in Figure 16, 17 and 18 the change in the accuracy of the verification set under different combinations of dropout rates and learning rates. But the effect of fine-tuning the pre-trained CNN model is still very poor. In addition, we also customize the CNN model to classify

TABLE 8. Comparison of the results of the RELIEF-SVM model and classifier.

Test / training (%)	LUSC-ASC	LUSC-SCLC	ASC-SCLC
SVM	73.91/100.00	59.57/93.96	75.33/97.61
LASSO-SVM	64.35/90.59	65.22/66.67	50.00/82.54
RELIEF-SVM	73.91/100.00	83.91/95.21	73.67/98.51
LDA-SVM	65.22/86.08	65.22/66.67	50.00/73.13
BP	73.91/82.35	56.52/95.83	86.67/100.00
KNN	60.87/92.16	60.87/75.00	76.67/82.09
Decision Tree	69.57/100.00	65.22/97.92	83.33/100.00
NB	69.57/78.43	56.52/77.08	76.67/77.61

Custom model architecture

Input
Conv 1 -64
Maxpool
Conv 2 -64
Maxpool
Conv 3 -128
Maxpool
FC -128
Dropout
FC -64
Dropout
FC -3
Softmax
Classification Layer

FIGURE 11. Custom model architecture.

lung cancer subtypes, and the experimental results are still not ideal. Therefore, when the availability of the original medical data set is limited, the input information still comes from a small number of original images, so the input information is still highly relevant. We cannot generate new information, but can only mix existing information, and cannot completely eliminate the impact of overfitting. The trained CNNs model, fine-tuning pre-trained CNNs structure, and custom model show poor classification performance when facing a small number of medical images [83], [84]. On the contrary, the classification effect of Relief-SVM model is better, which is more meets our expected effect.

IV. DISCUSSIONS

Histopathological examination is the gold standard for qualitative and clinical staging of lung tumors. Based on the local microvascular morphology, nucleus and cytoplasm on the histopathological images, doctors can diagnose tumor lesions. But for inexperienced or less experienced doctors, this is still a very difficult task. This study uses traditional feature extraction methods, trained CNNs, fine-tuning pre-trained CNNs models, and custom models to divide lung histopathological images into LUSC, ASC and SCLC.

Firstly, we extracted 265-dimensional features using seven handcrafted texture extraction algorithms, selected LDA feature dimensionality reduction algorithm, Relief and LASSO feature selection algorithms, and used a variety of mainstream classification algorithms for horizontal comparison, which proved that the Releif-SVM model has the best classification performance. The above algorithms all have their own advantages, but the LDA algorithm is more suitable for data with Gaussian distribution, so the performance of LDA-SVM in this research is poor [85]–[87]. We also selected four typical CNN models, and trained VGG16, ResNet50, DenseNet201, and InceptionV3 under the optimization of SGD algorithm and Nadam algorithm. In the experiment of
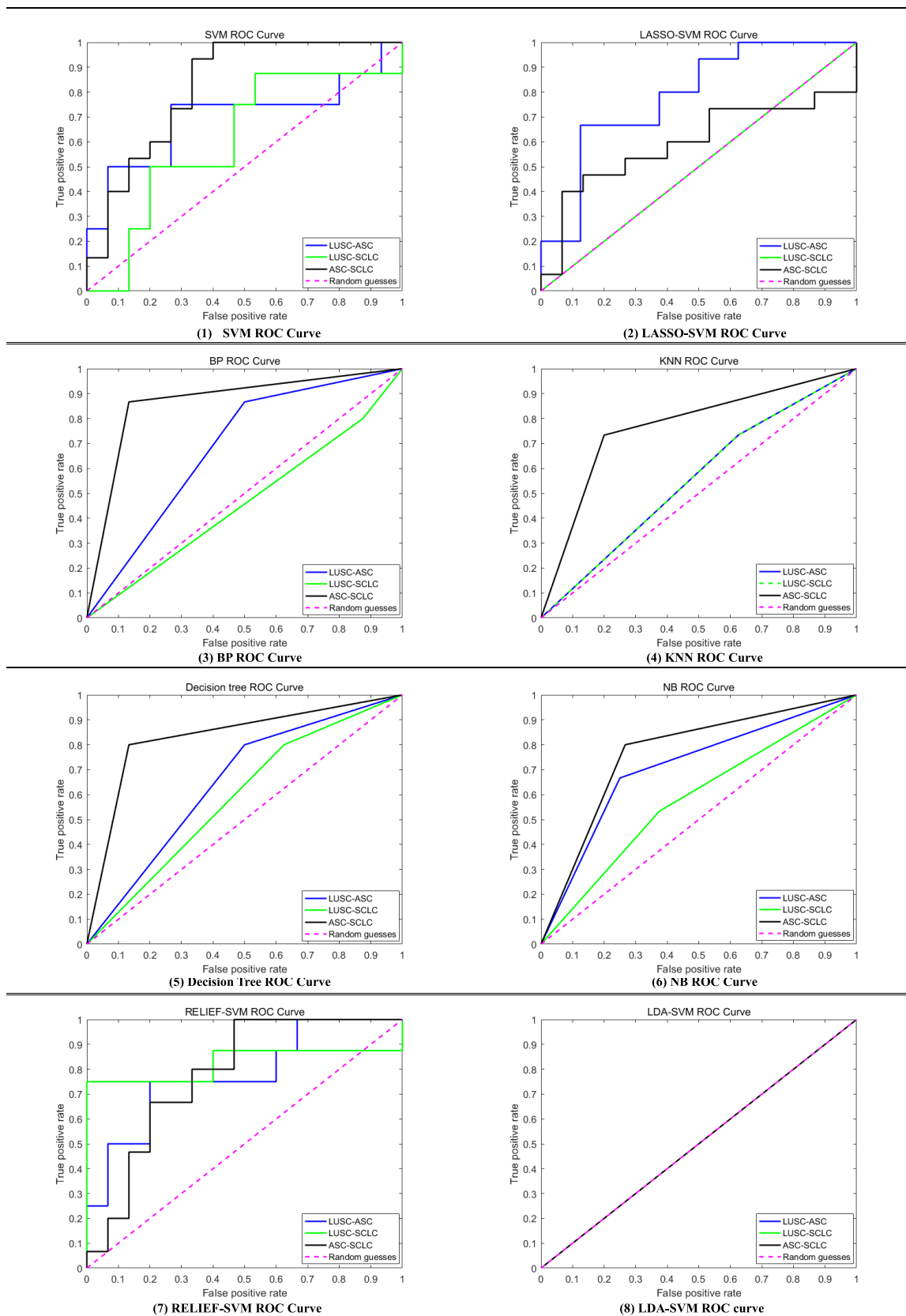
TABLE 9. ROC curve.

(1) SVM ROC Curve

(2) LASSO-SVM ROC Curve

(3) BP ROC Curve

(4) KNN ROC Curve

(5) Decision Tree ROC Curve

(6) NB ROC Curve

(7) RELIEF-SVM ROC Curve

(8) LDA-SVM ROC curve

TABLE 10. The AUC of eight classifiers.

AUC	LUSC-ASC	LUSC-SCLC	ASC-SCLC
SVM	0.7000	0.6083	0.8222
LASSO-SVM	0.7833	0.5000	0.5956
RELIEF-SVM	0.7750	0.8250	0.7778
LDA-SVM	0.5000	0.5000	0.5000
BP	0.6833	0.4625	0.8667
KNN	0.5542	0.5542	0.7667
Decision Tree	0.6500	0.5875	0.8333
NB	0.7083	0.5792	0.7667

TABLE 11. Display the number of images used for CNN model.

Types	Training		Validation	
	Training set	Augmented image	Validation set	Augmented image
LUSC	16	1600	8	800
ASC	35	3500	15	1500
SCLC	32	3200	15	1500

TABLE 12. The primary experimental parameter of the architecture model.

Architecture model	Parameter	Parameter value	Parameter	Parameter value
VGG16	Simulation environment	64-bit Windows 10 operating system	Software	PyCharm
	Image dimensions	224×224	Loss function	Categorical Cross-entropy
	Batch size	32	Epoch	50
	Optimization algorithm	Stochastic gradient descent	Learning rate	0.01
	Optimization algorithm	Nadam	Learning rate	0.002
ResNet50	Image dimensions	224×224	Loss function	Categorical Cross-entropy
	Batch size	32	Epoch	50
	Optimization algorithm	Stochastic gradient descent	Learning rate	0.01
	Optimization algorithm	Nadam	Learning rate	0.002
InceptionV3	Image dimensions	299×299	Loss function	Categorical Cross-entropy
	Batch size	32	Epoch	50
	Optimization algorithm	Stochastic gradient descent	Learning rate	0.01
	Optimization algorithm	Nadam	Learning rate	0.002
DenseNet201	Image dimensions	224×224	Loss function	Categorical Cross-entropy
	Batch size	32	Epoch	50
	Optimization algorithm	Stochastic gradient descent	Learning rate	0.01
	Optimization algorithm	Nadam	Learning rate	0.002

trained the CNNs model, the performance of ResNet50 and DensNet201 under the Nadam optimization algorithm is better than that of VGG16 and InceptionV3. On this basis, we apply the Nadam optimization algorithm to fine-tune the pre-trained CNNs model and custom models to try to find the best classification model. However, compared with the classification performance of Relief-SVM on the histopathological images of lung cancer subtypes, the DenseNet201 model still has a lower accuracy. In addition, we explored the collocation and combination of different dropout rates and learning rates under different experimental backgrounds. The results of Fig. 16, 17, 18 shown that there is still an

Trained CNNs
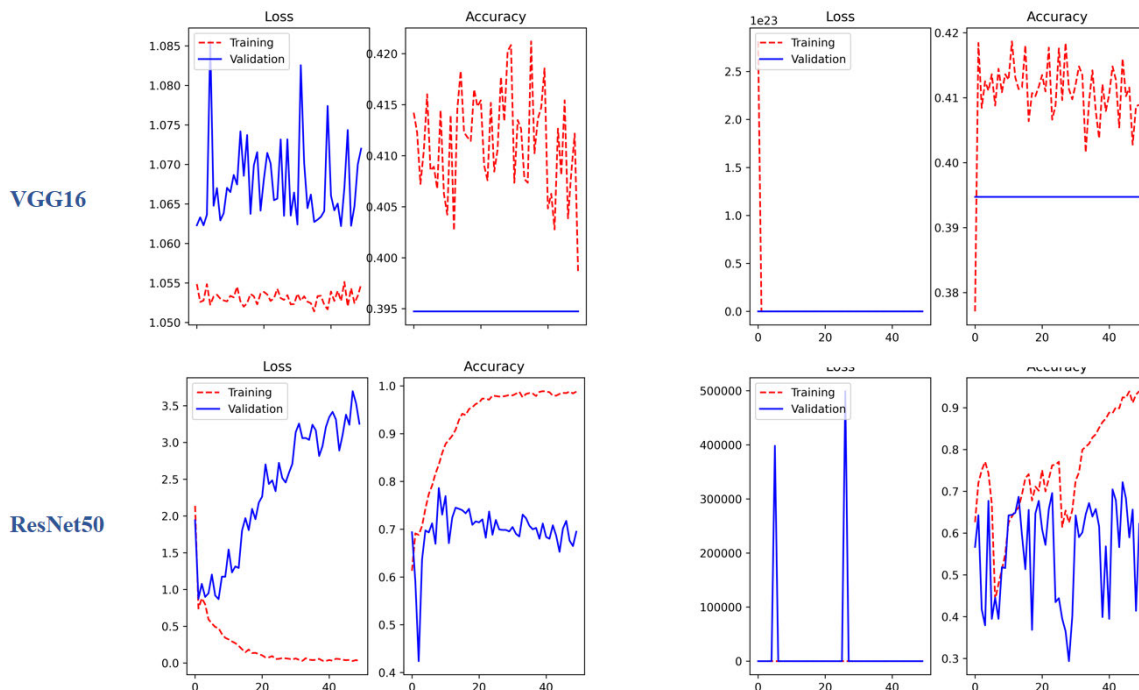
FIGURE 12. Visual picture of the trained CNN model.
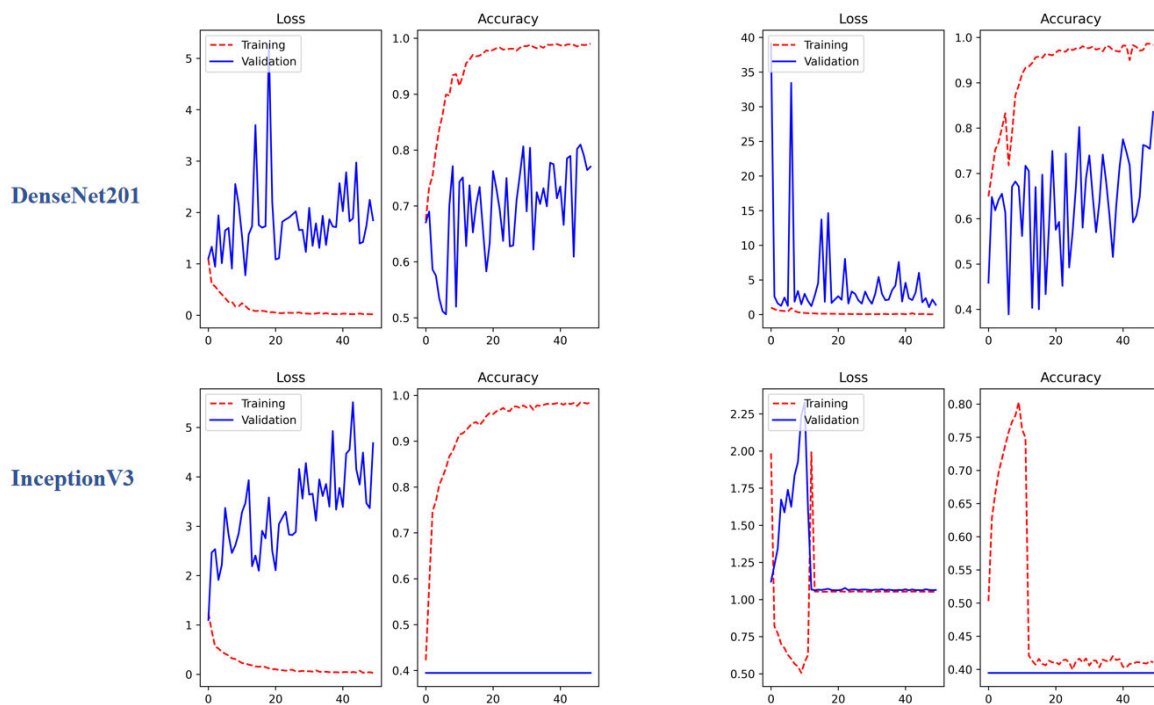
Trained CNNs

FIGURE 13. Visual picture of the trained CNN model.

overfitting phenomenon. However, the good news is that trained DenseNet201 shows good performance in these three combinations.

The results show that the Relief-SVM model built by machine learning is more suitable for the classification of lung histopathology images with a small data set. It means
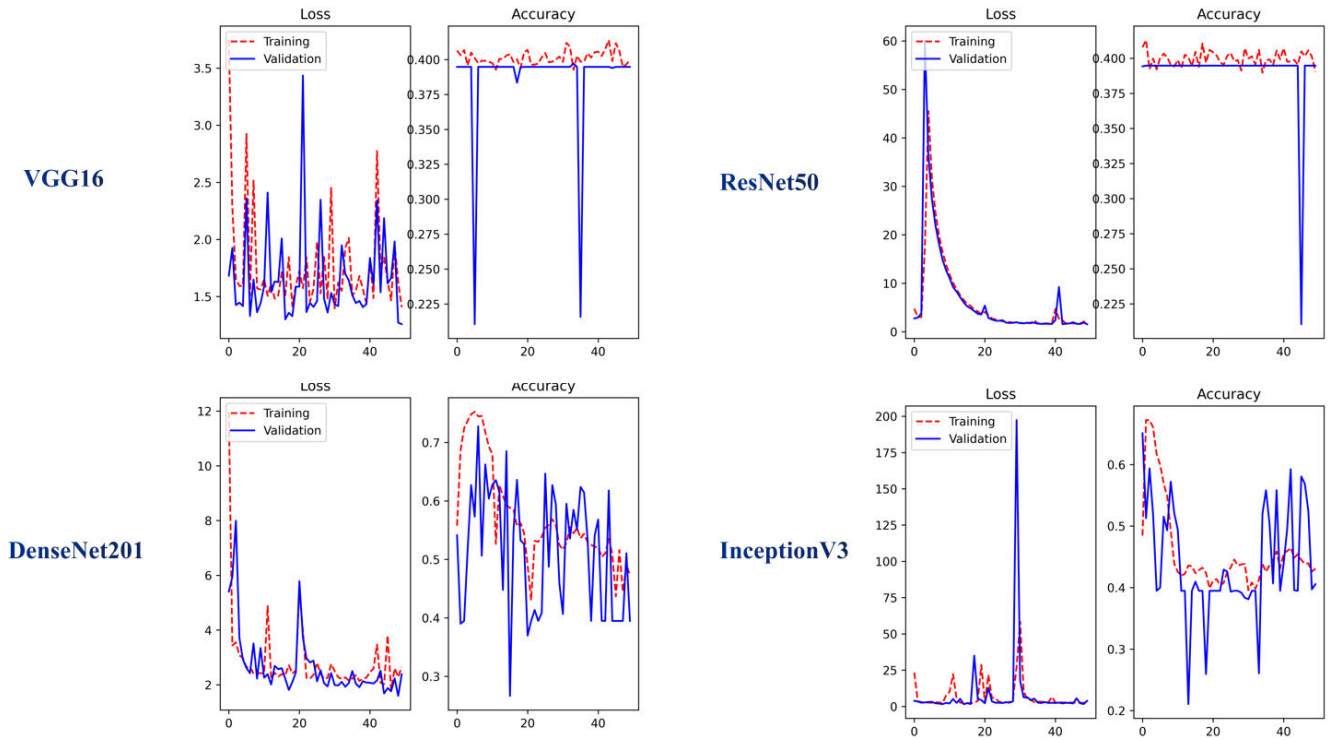
Fine-tuning Pre-trained CNNs-Nadam-optimized

FIGURE 14. Visual picture of the fine-tuning pre-trained CNNs model.

Custom model
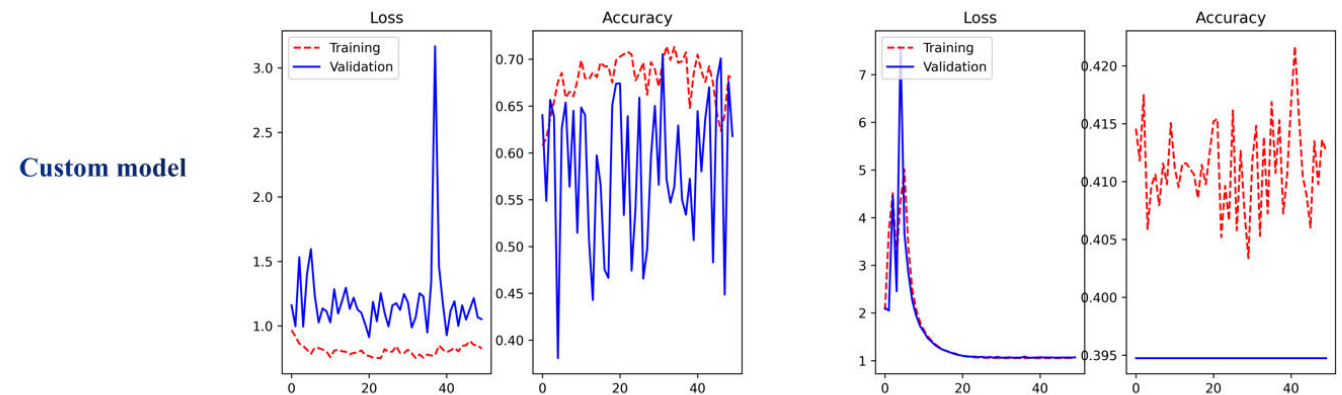
FIGURE 15. Visual picture of the custom model.

that for a limited number of data sets, especially small medical data sets, machine learning methods are often more effective for classification problems [84], [88]–[90]. In addition, the CNN features extracted from the DenseNet201 structure also prove from the side that it also plays a role in the task of classification of lung cancer subtypes.

According to the survey, the patch-based image classification method has been widely used in various histopathological images, which to a certain extent alleviates the problem of difficult and small number of medical image collection. However, due to the complexity of the texture of the tissue image and the time-consuming process of image annotation, there are currently few medical image annotation

Dropout rate = 0.5

FIGURE 16. The dropout rate is 0.5 with Nadam optimizer (lr = 0.1).
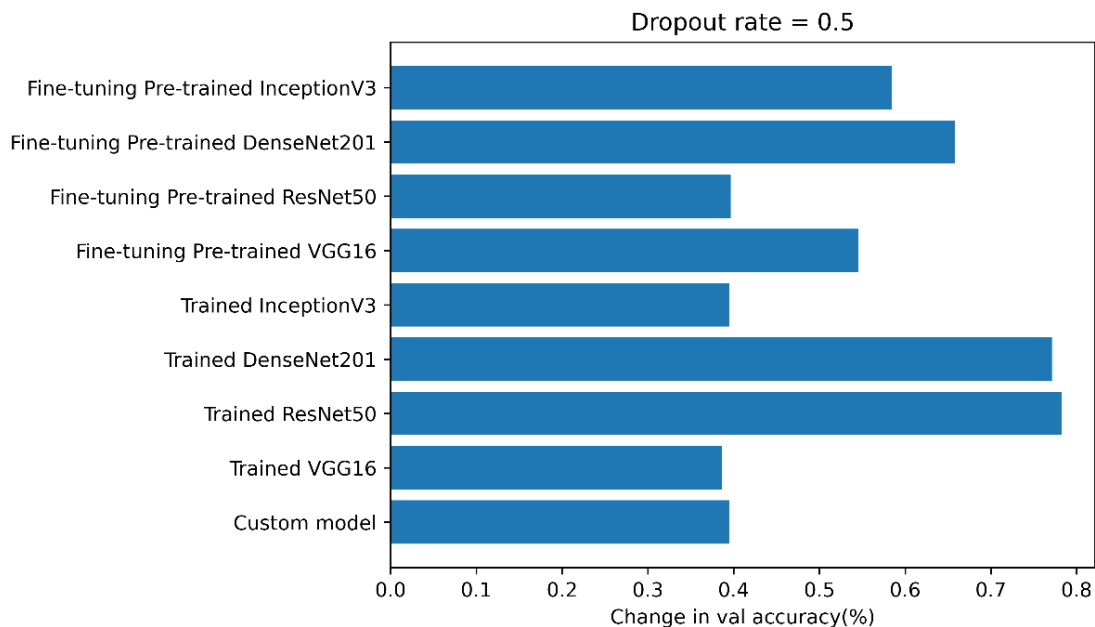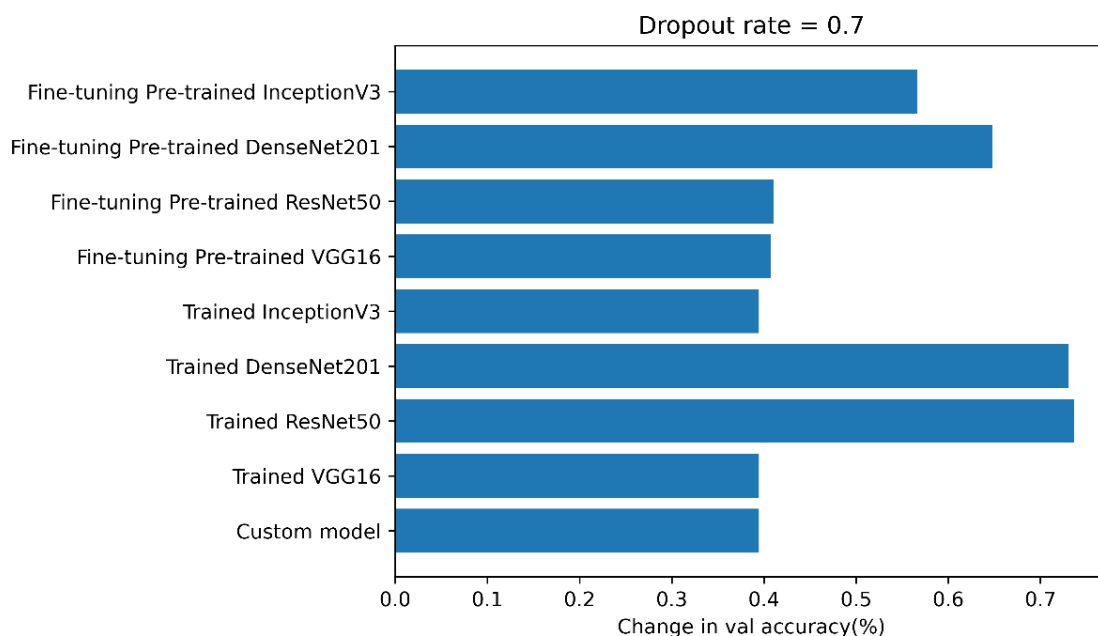
Dropout rate = 0.7

FIGURE 17. The dropout rate is 0.7 with Nadam optimizer (lr = 0.001).

data with a large number and comprehensive labeling data. In addition, since only image-level tags exist in the data set, how to distinguish between patch-level tags and image-level tags is a huge challenge for patch-based classification methods.

It is worth noting that in previous work, there have been studies on classification of lung cancer subtypes based on computer-aided diagnosis. However, according to research, there has not been any study that includes histopathological

images of ASC for automatic classification. Therefore, this study included ASC samples for the first time. The Relief-SVM model used in this paper can achieve an overall accuracy of more than 70%. This is a satisfactory result, which can help the application of cytology in the diagnosis of lung cancer. In addition, we also summarized the research related to this article as shown in Table 13 [21], [91]–[96].

In summary, the classification capabilities of the Relief-SVM model is superior to the trained CNNs, fine-tuned

TABLE 13. Comparison with previous work.

Paper number	Topic	Classification	Accuracy	Advantages and disadvantages
[21]	Automated morphological classification of lung cancer subtypes using H&E tissue images	Adenocarcinoma and squamous carcinoma	The accuracy rate is as high as 90% on tissue microarray cores and full tissue sections.	This study is the first to prove the importance of tissue pattern measurement, which is more conducive to distinguishing SC and non-SC histotypes. However, this study has the problem of incomplete subjects.
[91]	Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks	Adenocarcinoma, squamous cell carcinoma, and small cell carcinoma	The total correct rate by augmented images achieved 71.1%.	This research is the first to carry out an exploratory study using cytological images. However, this study has the problem of incomplete subjects.
[92]	Deep Learning for the Classification of Small-Cell and Non-Small-Cell Lung Cancer	Pulmonary adenocarcinoma, pulmonary squamous cell carcinoma, and small-cell lung cancer	Patient-based CNN classification results was 100% in the test set after the application of strict quality control (QC).	Skeletal muscle samples are included in the analysis in order to check the plausibility of the results. However, this study did not explore the classification performance achieved by traditional feature extraction methods, and the ASC samples were not included in the experimental subjects.
[93]	Multi-subtype classification model for non-small cell lung cancer based on radiomics: SLS model	squamous cell carcinoma and adenocarcinoma, and the other two subtypes of non-small cell lung cancer (NSCLC): large cell carcinoma and not otherwise specified	The classification accuracy of 0.86 on the test set.	The study was included in the LUNG1 and LUNG3 datasets, which the data we used was more diverse and complex. However, more attention should be paid to the potential impact of different imaging setups in the test set.
[94]	Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer	Squamous cell carcinoma and adenocarcinoma	In the validation cohort, the radiomic signature showed an AUC of 0.893	The constructed radiomic signature has very good classification results for lung adenocarcinoma and squamous cell carcinoma. However, small sample size and single-center cohort problems still exist.
[95]	Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)	Normal、Abnormal (The patient has lung cancer)	The classification accuracy in test set of 86.6%.	The study was included in the Kaggles Data Science Bowl (DSB) 2017 and Lung Nodule Analysis 2016 (LUNA16) Challenge, which is very exploratory. We hope that subsequent studies can be extended to 3D images of other cancers.
[96]	A comparison study on the effect of false positive reduction in deep learning based detection for juxtapleural lung nodules: Cnn vs dnn	Lung Nodule Detection	The CNN based classifier achieved sensitivity of 0.82 for lung juxtapleural detection.	This study applied two types of deep learning networks to reduce the false positive rates for the detected nodules. However, small sample size still exit.

pre-trained CNNs model, and custom models. However, our work also has some limitations. On the one hand, this study collected a limited number of training set images from the initial stage to establish a more accurate classification accuracy, and did not try other numbers of training set images. If medical image data from multiple centers with different classifications can be used for research, the diversity of the data will be further enhanced and the classification performance of the model will be improved. In addition, due to the general influence of external factors, the data set is susceptible to color

and noise. The use of the above image preprocessing methods has certain limitations and may potentially affect the test results. Therefore, when we face more abundant data sources in the follow-up research, we should pay more attention to the preparation and collection of histopathological images. On the other hand, if a larger amount of lung histopathology image data can be included for research, we will obtain more complete disease information, which will further improve the generalization of the classification model. Therefore, the histological subtype classification of lung cancer based
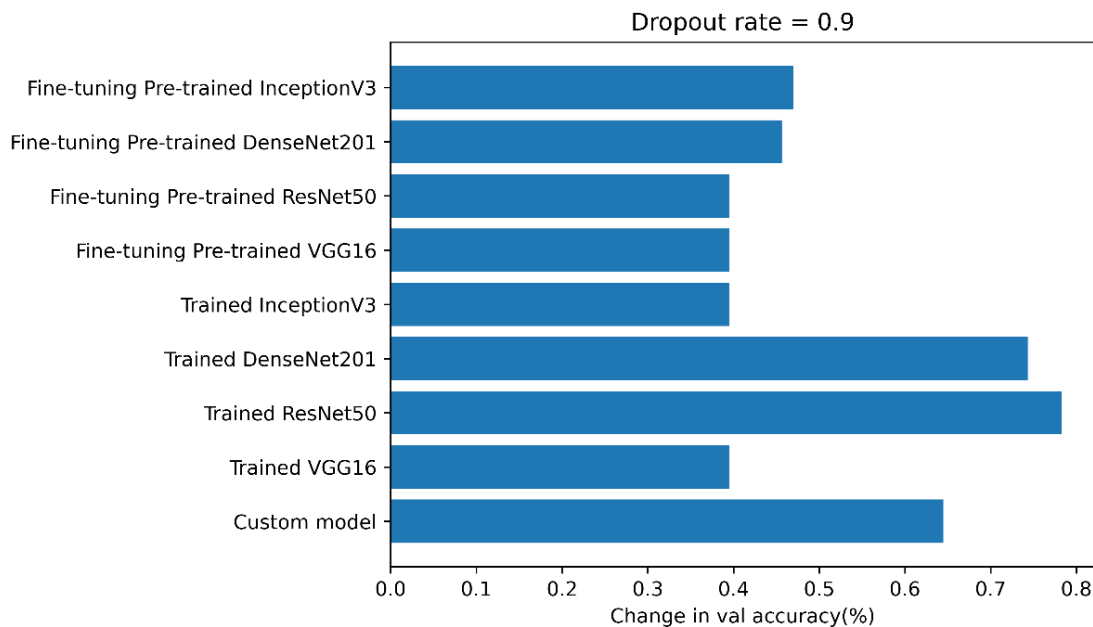
FIGURE 18. The dropout rate is 0.9 with Nadam optimizer (lr = 0.00001).

on histopathological images is worthy of our further study to provide clinicians with more objective auxiliary diagnosis results.

V. CONCLUSION

This paper was the first to apply the Relief-SVM algorithm to the classification of LC histopathological images. We included 121 histopathological images of LUSC, ASC and SCLC. The experimental results show that Relief-SVM has the best classification performance in distinguishing lung cancer subtypes, which provides a favorable guidance for the classification of lung histopathological images. And it also provides a correct direction for the classification of medical images with complex texture tissue. In addition, this paper analyzes and explores the classification effects under the traditional manual method of extracting features, trained the CNNs model, fine-tuning pre-trained CNNs models, and custom models to find which classification model is more suitable for the classification of different subtypes of lung cancer.

The experimental results show that the Relief-SVM classification model achieves best classification performance regardless of whether it is compared with CNN models which are trained directly by using the histopathology image data set or compared with the fine-tuning pre-trained CNNs models and custom models. In other words, when faced with a small number of medical images, traditional manual methods have shown great potential in the classification of lung histopathological images. However, the CNN models which are trained directly by using the histopathology image data set and the fine-tuning pre-trained CNNs models and custom models CNN model show weak classification performance.

Notably, our study has a considerable guiding function for doctors with less experience, which can provide objective reference results and relieve the heavy work pressure of doctors. It is feasible and practical to a certain extent. We will continue to collect more samples from different central institutions and build a lung histopathological image database for lung diseases in the future.

VI. CONFLICTS OF INTEREST

The authors have no relevant financial interests in this article and no potential conflicts of interest to disclose.

ACKNOWLEDGMENT

REFERENCES

[1] W. Cao, R. Wu, G. Cao, and Z. He, "A comprehensive review of computer-aided diagnosis of pulmonary nodules based on computed tomography scans," *IEEE Access*, vol. 8, pp. 154007–154023, 2020.

[2] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai, and H. Hu, "Automated lung nodule detection and classification using deep learning combined with multiple strategies," *Sensors*, vol. 19, no. 17, p. 3722, Aug. 2019.

[3] X. Zhao, S. Qi, B. Zhang, H. Ma, W. Qian, Y. Yao, and J. Sun, "Deep CNN models for pulmonary nodule classification: Model modification, model integration, and transfer learning," *J. X-Ray Sci. Technol.*, vol. 27, no. 4, pp. 615–629, Sep. 2019.

[4] M. Kriegsmann, C. Haag, C. A. Weis, G. Steinbuss, and K. Kriegsmann, "Deep learning for the classification of small-cell and non-small-cell lung cancer," *Cancers*, vol. 12, no. 6, p. 1604, 2020.

[5] P. Ocampo, A. Moreira, N. Coudray, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyo, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning," *J. Thoracic Oncol.*, vol. 13, no. 10, p. S562, Oct. 2018.

[6] X. Liu, C. Wang, J. Bai, and G. Liao, "Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images," *Neurocomputing*, vol. 392, pp. 253–267, Jun. 2020.

[7] F. Kanavati, G. Toyokawa, S. Momosaki, M. Rambeau, Y. Kozuma, F. Shoji, K. Yamazaki, S. Takeo, O. Iizuka, and M. Tsuneki, "Weakly-supervised learning for lung carcinoma classification using deep learning," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, Jun. 2020.

[8] Y. Han, Y. Ma, Z. Wu, F. Zhang, D. Zheng, X. Liu, L. Tao, Z. Liang, Z. Yang, X. Li, J. Huang, and X. Guo, "Histologic subtype classification of non-small cell lung cancer using PET/CT images," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 48, no. 2, pp. 350–360, Feb. 2021.

[9] S. Lian, Y. Huang, H. Yang, and H. Zhao, "Serum carbohydrate antigen 12-5 level enhances the prognostic value in primary adenosquamous carcinoma of the lung: A two-institutional experience," *Interact. Cardio-Vascular Thoracic Surg.*, vol. 22, no. 4, pp. 419–424, Apr. 2016.

[10] H. Maeda, A. Matsumura, T. Kawabata, T. Suito, O. Kawashima, T. Watanabe, K. Okabayashi, and I. Kubota, "Adenosquamous carcinoma of the lung: Surgical results as compared with squamous cell and adenocarcinoma cases," *Eur. J. Cardio-Thoracic Surg.*, vol. 41, no. 2, pp. 357–361, Feb. 2012.

[11] T. Miki, S. Yano, M. Hanibuchi, and S. Sone, "Bone metastasis model with multiorgan dissemination of human small-cell lung cancer (SBC-5) cells in natural killer cell-depleted SCID mice," *Oncol. Res. Featuring Preclinical Clin. Cancer Therapeutics*, vol. 12, no. 5, pp. 209–217, May 2001.

[12] R. L. Draelos, D. Dov, M. A. Mazurowski, J. Y. Lo, R. Henao, G. D. Rubin, and L. Carin, "Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101857.

[13] Y. Kang, Y. Fang, and X. Lai, "Automatic detection of diabetic retinopathy with statistical method and Bayesian classifier," *J. Med. Imag. Health Informat.*, vol. 10, no. 5, pp. 1225–1233, May 2020.

[14] W. Wang and L. Zhang, "Magnetic resonance imaging manifestations of brain metastases in patients with lung cancer," *J. Med. Imag. Health Informat.*, vol. 10, no. 12, pp. 2985–2988, Dec. 2020.

[15] M. Szabó, A. Bozó, K. Darvas, S. Soós, M. Özse, and Z. D. Iványi, "The role of ultrasonographic lung aeration score in the prediction of postoperative pulmonary complications: An observational study," *BMC Anesthesiology*, vol. 21, no. 1, pp. 1–10, Dec. 2021.

[16] C. T. Sari and C. Gunduz-Demir, "Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1139–1149, May 2019.

[17] P. J. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, and P. Honeine, "Multiple instance learning for histopathological breast cancer image classification," *Expert Syst. Appl.*, vol. 117, pp. 103–111, Mar. 2019.

[18] M. Li, X. Nie, Y. Reheman, P. Huang, S. Zhang, Y. Yuan, C. Chen, Z. Yan, C. Chen, X. Lv, and W. Han, "Computer-aided diagnosis and staging of pancreatic cancer based on CT images," *IEEE Access*, vol. 8, pp. 141705–141718, 2020.

[19] S. Pandey, P. R. Singh, and J. Tian, "An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101782.

[20] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation for the training of deep neural networks," *Neural Comput. Appl.*, vol. 32, no. 19, pp. 15503–15531, Oct. 2020.

[21] C.-W. Wang and C.-P. Yu, "Automated morphological classification of lung cancer subtypes using H&E tissue images," *Mach. Vis. Appl.*, vol. 24, no. 7, pp. 1383–1391, Oct. 2013.

[22] S. Boumaraf, X. Liu, Z. Zheng, X. Ma, and C. Ferkous, "A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102192.

[23] A. R. Fernandez, M. E. B. Guardia, M. G. Rio, C. R. Font, A. S. P. Ramos, J. M. L. Elvira, and V. P. Muriel, "Staging of non-small cell lung cancer. Diagnosis efficacy of structural (CT) and functional (FDG-PET) imaging methods," *Revista Clinica Espanola*, vol. 207, no. 11, pp. 541–547, Dec. 2007.

[24] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 1868–1873.

[25] F. Ciompi, O. Geessink, B. E. Bejnordi, G. S. de Souza, A. Baidoshvili, G. Litjens, B. van Ginneken, I. Nagtegaal, and J. van der Laak, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 160–163.

[26] C. Sun, A. Xu, D. Liu, Z. Xiong, F. Zhao, and W. Ding, "Deep learning-based classification of liver cancer histopathology images using only global labels," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1643–1651, Jun. 2020.

[27] M. Shahmoradi, M. Lashgari, H. Rabbani, J. Qin, and M. Swain, "A comparative study of new and current methods for dental micro-CT image denoising," *Dentomaxillofacial Radiol.*, vol. 45, no. 3, Mar. 2016, Art. no. 20150302.

[28] Ş. Öztürk and B. Akdemir, "HIC-Net: A deep convolutional neural network model for classification of histopathological breast images," *Comput. Electr. Eng.*, vol. 76, pp. 299–310, Jun. 2019.

[29] S. Alinsaif and J. Lang, "Texture features in the shearlet domain for histopathological image classification," *BMC Med. Informat. Decis. Making*, vol. 20, no. S14, pp. 1–19, Dec. 2020.

[30] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proc. 12th Int. Conf. Pattern Recognit.*, Oct. 1994, pp. 582–585.

[31] P. Huang, S. Zhang, M. Li, J. Wang, C. Ma, B. Wang, and X. Lv, "Classification of cervical biopsy images based on LASSO and EL-SVM," *IEEE Access*, vol. 8, pp. 24219–24228, 2020.

[32] P. Parvathaneni, I. Lyu, Y. Huo, B. P. Rogers, K. G. Schilling, V. Nath, J. A. Blaber, A. E. Hainline, A. W. Anderson, N. D. Woodward, and B. A. Landman, "Improved gray matter surface based spatial statistics in neuroimaging studies," *Magn. Reson. Imag.*, vol. 61, pp. 285–295, Sep. 2019.

[33] B. Tu, W. Kuang, G. Zhao, D. He, Z. Liao, and W. Ma, "Hyperspectral image classification by combining local binary pattern and joint sparse representation," *Int. J. Remote Sens.*, vol. 40, no. 24, pp. 9484–9500, Dec. 2019.

[34] A. Nithya, A. Appathurai, N. Venkatadri, D. R. Ramji, and C. Anna Palagan, "Kidney disease detection and segmentation using artificial neural network and multi-kernel K-means clustering for ultrasound images," *Measurement*, vol. 149, Jan. 2020, Art. no. 106952.

[35] H. Liu, H. Jiang, B. Xia, and D. Yi, "The research of feature extraction method of liver pathological image based on multispatial mapping and statistical properties," *Comput. Math. Methods Med.*, vol. 2016, pp. 1–8, Jan. 2016.

[36] C. Liu, H. Zhang, Y. Pan, F. Huang, and S. Xia, "Towards MIB-1 and p53 detection in glioma magnetic resonance image: A novel computational image analysis method," *Phys. Med. Biol.*, vol. 57, no. 24, p. 8393, 2012.

[37] Y. Ren, J. Yang, Q. Zhang, and Z. Guo, "Ship recognition based on Hu invariant moments and convolutional neural network for video surveillance," *Multimedia Tools Appl.*, vol. 80, no. 1, pp. 1343–1373, Jan. 2021.

[38] A. Hjouji, J. EL-Mekkaoui, and M. Jourhmane, "Rotation scaling and translation invariants by a remediation of Hu's invariant moments," *Multimedia Tools Appl.*, vol. 79, nos. 19–20, pp. 14225–14263, May 2020.

[39] C. I. Christodoulou, E. Kyriacou, M. S. Pattichis, C. S. Pattichis, and A. Nicolaides, "A comparative study of morphological and other texture features for the characterization of atherosclerotic carotid plaques," in *Proc. 10th Int. Conf. Comput. Anal. Images Patterns (CAIP)*, Groningen, The Netherlands. Berlin, Germany: Springer, Aug. 2003, pp. 503–511.

[40] P.-Y. Liu and Z.-M. Li, "A feature extraction method based on local binary pattern preprocessing and wavelet transform," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 34, no. 13, Dec. 2020, Art. no. 2050030.

[41] I. Hedhli, G. Moser, S. B. Serpico, and J. Zerubia, "Classification of multisensor and multiresolution remote sensing images through hierarchical Markov random fields," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2448–2452, Dec. 2017.

[42] X. Huang, M. Chen, P. Liu, and Y. Du, "Texture feature-based classification on transrectal ultrasound image for prostatic cancer detection," *Comput. Math. Methods Med.*, vol. 2020, pp. 1–9, Oct. 2020.

[43] N. Q. K. Le, "Fertility-GRU: Identifying fertility-related proteins by incorporating deep-gated recurrent units and original position-specific scoring matrix profiles," *J. Proteome Res.*, vol. 18, no. 9, pp. 3503–3511, Sep. 2019.

[44] N.-Q.-K. Le and B. P. Nguyen, "Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Aug. 1, 2020, doi: 10.1109/TCBB.2019.2932416.

[45] Y. Shen, D. Cao, K. Ruddy, and L. F. Teixeira de Moraes, "Near real-time hydraulic fracturing event recognition using deep learning methods," *SPE Drilling Completion*, vol. 35, no. 3, pp. 478–489, Sep. 2020.

[46] G. Pengcheng, G. Gang, W. Jiangqin, and W. Baogang, "Chinese calligraphic style representation for recognition," *Int. J. Document Anal. Recognit.*, vol. 20, no. 1, pp. 59–68, Mar. 2017.

[47] A. Sedik *et al.*, "Efficient deep learning approach for augmented detection of Coronavirus disease," *Neural Comput. Appl.*, Jan. 2021, doi: 10.1007/s00521-020-05410-8.

[48] A. Sedik, A. M. Iliyasu, B. Abd El-Rahiem, M. E. Abdel Samea, A. Abdel-Raheem, M. Hammad, J. Peng, F. E. Abd El-Samie, and A. A. Abd El-Latif, "Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections," *Viruses*, vol. 12, no. 7, p. 769, Jul. 2020.

[49] M. Hammad *et al.*, "Myocardial infarction detection based on deep neural network on imbalanced data," *Multimedia Syst.*, Jan. 2021, doi: 10.1007/s00530-020-00728-8.

[50] M. Hammad, P. Pławiak, K. Wang, and U. R. Acharya, "ResNet—Attention model for human authentication using ECG signals," *Expert Syst.*, 2020, Art. no. e12547, doi: 10.1111/exsy.12547.

[51] P. K. Shukla, R. K. Chaurasiya, and S. Verma, "Performance improvement of P300-based home appliances control classification using convolution neural network," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102220.

[52] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Ser. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[53] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, "Feature selection for gene expression using model-based entropy," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 1, pp. 25–36, Jan. 2010.

[54] H. Liu, H. Motoda, and L. Yu, "A selective sampling approach to active feature selection," *Artif. Intell.*, vol. 159, nos. 1–2, pp. 49–74, Nov. 2004.

[55] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Informat.*, vol. 85, pp. 189–203, Sep. 2018.

[56] A. M. D. Younus and B.-S. Yang, "Intelligent fault diagnosis of rotating machinery using infrared thermal image," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 2082–2091, Feb. 2012.

[57] W. Zhang, Y. Wu, Z. Lei, and Z. Liu, "Feature selection for SNP data based on relief-SVM," in *Proc. 4th Int. Conf. Biomed. Eng. Informat. (BMEI)*, Oct. 2011, pp. 1720–1723.

[58] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. AAAI*, 1992, pp. 129–134.

[59] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. 9th Int. Workshop Mach. Learn. (ML)*. Scotland, U.K.: Morgan Kaufmann, Jul. 1992, doi: 10.1016/B978-1-55860-247-2.50037-1.

[60] R. J. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Stat. Soc., Ser. B, Methodol.*, vol. 73, no. 1, pp. 273–282, 1996.

[61] W. N. Venables and B. D. Ripley, *Modern Applied Statistics With S*. Springer, 2010.

[62] Z. Tan, A. Jamdagni, X. He, and P. Nanda, "Network intrusion detection based on LDA for payload feature selection," in *Proc. IEEE Globecom Workshops*, Dec. 2010, pp. 1545–1549.

[63] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, May 2010.

[64] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu, "SGL-SVM: A novel method for tumor classification via support vector machine with sparse group lasso," *J. Theor. Biol.*, vol. 486, Feb. 2020, Art. no. 110098.

[65] J. Liu, J. Wang, Z. Tang, B. Hu, F.-X. Wu, and Y. Pan, "Improving Alzheimer's disease classification by combining multiple measures," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 5, pp. 1649–1659, Sep. 2018.

[66] J. Liu, M. Li, W. Lan, F.-X. Wu, Y. Pan, and J. Wang, "Classification of Alzheimer's disease using whole brain hierarchical network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 2, pp. 624–632, Mar. 2018.

[67] W. Li, Z. Li, H. Zhang, Y. Wang, H. Chen, and L. Xiong, "Acupoint catgut embedding for insomnia: A meta-analysis of randomized controlled trials," *Evidence-Based Complementary Alternative Med.*, vol. 2020, pp. 1–14, Nov. 2020.

[68] C. Reynders-Frederix, P. Reynders-Frederix, I. Bernardo, and B. Mihai, "Development of an electronic assistive walking device," *Romanian Biotechnol. Lett.*, vol. 25, no. 5, pp. 1992–1997, Sep. 2020.

[69] J. Liu, M. Li, Y. Pan, F.-X. Wu, X. Chen, and J. Wang, "Classification of schizophrenia based on individual hierarchical brain networks constructed from structural MRI images," *IEEE Trans. Nanobiosci.*, vol. 16, no. 7, pp. 600–608, Oct. 2017.

[70] P. Huang, X. Tan, C. Chen, X. Lv, and Y. Li, "AF-SENet: Classification of cancer in cervical tissue pathological images based on fusing deep convolution features," *Sensors*, vol. 21, no. 1, p. 122, Dec. 2020.

[71] L. Rampasek and A. Goldenberg, "TensorFlow: Biology's gateway to deep learning?" *Cell Syst.*, vol. 2, no. 1, pp. 12–14, Jan. 2016.

[72] K. Sriporn, C.-F. Tsai, C.-E. Tsai, and P. Wang, "Analyzing lung disease using highly effective deep learning techniques," *Healthcare*, vol. 8, no. 2, p. 107, Apr. 2020.

[73] M. Toğaçar, K. B. Özkurt, B. Ergen, and Z. Cömert, "BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer," *Phys. A, Stat. Mech. Appl.*, vol. 545, May 2020, Art. no. 123592.

[74] T. Kausar, M. Wang, M. Idrees, and Y. Lu, "HWDCNN: Multi-class recognition in breast histopathology with Haar wavelet decomposed image based convolution neural network," *Biocybern. Biomed. Eng.*, vol. 39, no. 4, pp. 967–982, Oct. 2019.

[75] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[76] I. Banerjee, A. Crawley, M. Bhethanabotla, H. E. Daldrup-Link, and D. L. Rubin, "Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma," *Computerized Med. Imag. Graph.*, vol. 65, pp. 167–175, Apr. 2018.

[77] P. Maduskar, R. H. M. M. Philipsen, J. Melendez, E. Scholten, D. Chanda, H. Ayles, C. I. Sánchez, and B. van Ginneken, "Automatic detection of pleural effusion in chest radiographs," *Med. Image Anal.*, vol. 28, pp. 22–32, Feb. 2016.

[78] R. Sivaramakrishnan, S. Antani, S. Candemir, Z. Xue, G. Thoma, P. Alderson, J. Abuya, and M. Kohli, "Comparing deep learning models for population screening using chest radiography," *Proc. SPIE*, vol. 10575, Feb. 2018, Art. no. 105751E.

[79] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[80] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI," *J. Magn. Reson. Imag.*, vol. 49, no. 4, pp. 939–954, 2018.

[81] A. Suzuki, M. Aoki, C. Miyagawa, K. Murakami, H. Takaya, Y. Kotani, H. Nakai, and N. Matsumura, "Differential diagnosis of uterine leiomyoma and uterine sarcoma using magnetic resonance images: A literature review," *Healthcare*, vol. 7, no. 4, p. 158, Dec. 2019.

[82] N. Noreen, S. Palaniappan, A. Qayyum, I. Ahmad, M. Imran, and M. Shoaib, "A deep learning model based on concatenation approach for the diagnosis of brain tumor," *IEEE Access*, vol. 8, pp. 55135–55144, 2020.

[83] D. Du, H. Feng, W. Lv, S. Ashrafinia, Q. Yuan, Q. Wang, W. Yang, Q. Feng, W. Chen, A. Rahmim, and L. Lu, "Machine learning methods for optimal radiomics-based differentiation between recurrence and inflammation: Application to nasopharyngeal carcinoma post-therapy PET/CT images," *Mol. Imag. Biol.*, vol. 22, no. 3, pp. 730–738, Jun. 2020.

[84] C. Chen, X. Guo, J. Wang, W. Guo, X. Ma, and J. Xu, "The diagnostic value of radiomics-based machine learning in predicting the grade of meningiomas using conventional magnetic resonance imaging: A preliminary study," *Frontiers Oncol.*, vol. 9, p. 1338, Dec. 2019.

[85] S. Yang, H. Xiong, K. Xu, L. Wang, J. Bian, and Z. Sun, "Improving covariance-regularized discriminant analysis for EHR-based predictive analytics of diseases," *Int. J. Speech Technol.*, vol. 51, no. 1, pp. 377–395, Jan. 2021.

[86] I. Cardoso, E. Almeida, H. Allende-Cid, A. Frery, R. Rangayyan, P. Azevedo-Marques, and H. Ramos, "Analysis of machine learning algorithms for diagnosis of diffuse lung diseases," *Methods Inf. Med.*, vol. 57, nos. 5–6, pp. 272–279, Nov. 2018.

[87] S. Zeng, C. Gao, X. Wang, L. Jiang, and D. Feng, "Multiple kernel-based discriminant analysis via support vectors for dimension reduction," *IEEE Access*, vol. 7, pp. 35418–35430, 2019.

[88] M. He, Z. Liu, Y. Lin, J. Wan, J. Li, K. Xu, Y. Wang, Z. Jin, J. Tian, and H. Xue, "Differentiation of atypical non-functional pancreatic neuroendocrine tumor and pancreatic ductal adenocarcinoma using CT based radiomics," *Eur. J. Radiol.*, vol. 117, pp. 102–111, Aug. 2019.

[89] T. Y. Rahman, L. B. Mahanta, H. Choudhury, A. K. Das, and J. D. Sarma, "Study of morphological and textural features for classification of oral squamous cell carcinoma by traditional machine learning techniques," *Cancer Rep.*, vol. 3, no. 6, Dec. 2020, Art. no. e1293.

[90] M. Nazari, I. Shiri, G. Hajianfar, N. Oveisi, H. Abdollahi, M. R. Deevband, M. Oveisi, and H. Zaidi, "Noninvasive fuhrman grading of clear cell renal cell carcinoma using computed tomography radiomic features and machine learning," *La Radiologia Medica*, vol. 125, no. 8, pp. 754–762, Aug. 2020.

[91] A. Teramoto, T. Tsukamoto, Y. Kiriyama, and H. Fujita, "Automated classification of lung cancer types from cytological images using deep convolutional neural networks," *BioMed Res. Int.*, vol. 2017, pp. 1–6, Aug. 2017.

[92] K. Kriegsmann, C. Haag, C. A. Weis, G. Steinbuss, and M. Kriegsmann, "Deep learning for the classification of small cell and non-small cell lung cancer," *Oncol. Res. And Treatment*, vol. 43, no.4, p. 144, Oct. 2020.

[93] J. Liu, J. Cui, F. Liu, Y. Yuan, F. Guo, and G. Zhang, "Multi-subtype classification model for non-small cell lung cancer based on radiomics: SLS model," *Med. Phys.*, vol. 46, no. 7, pp. 3091–3100, Jul. 2019.

[94] X. Zhu, D. Dong, Z. Chen, M. Fang, L. Zhang, J. Song, D. Yu, Y. Zang, Z. Liu, J. Shi, and J. Tian, "Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer," *Eur. Radiol.*, vol. 28, no. 7, pp. 2772–2778, Jul. 2018.

[95] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3D convolutional neural network (3D-CNN)," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 8, p. 409, 2017.

[96] J. Tan, Y. Huo, Z. Liang, and L. Li, "A comparison study on the effect of false positive reduction in deep learning based detection for juxtapleural lung nodules: CNN VS DNN," in *Proc. Symp. Modeling Simulation Med.*, Apr. 2017, pp. 1–8.
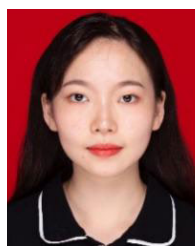
CHEN CHEN received the bachelor's degree from Xinjiang University, China, in June 2018, where he is currently pursuing the Ph.D. degree. His main research interest includes medical signal processing.

YUSHUAI YUAN received the bachelor's degree from the Tianjin University of Technology, China, in June 2018. He is currently pursuing the master's degree with Xinjiang University. His main research interest includes image processing.

SHUAILEI ZHANG received the bachelor's degree in communications engineering from Xinjiang University, China, in June 2019. She is currently pursuing the master's degree in information and communications engineering in China. She was recommended to be admitted to Xinjiang University, in September 2019. Her main research interests include image processing and bioinformatics.

ZIWEI YAN received the bachelor's degree from Xinjiang University, China, in June 2018, where she is currently pursuing the master's degree. Her main research interest includes medical signal processing.

MIN LI received the bachelor's degree in computer science and technology from Shandong Women's University, China, in 2019. She is currently pursuing the master's degree in software engineering with Xinjiang University. Her current research interest includes medical image processing.

CHENG CHEN received the bachelor's degree from Xinjiang University, China, in June 2018, where he is currently pursuing the master's degree. His main research interest includes bioinformatics.

XIAOJIAN MA has been working with the Department of Information Management, The Affiliated Cancer Hospital of Xinjiang Medical University, since 2010. For ten years, he has focused on the construction of hospital information management, has been committed to the research on the security and reliability of the core information system of large hospitals, and has rich practical experience in the construction and management of hospital information.

FANGFANG CHEN received the bachelor's degree from the College of Information Science and Engineering, Xinjiang University, China, in 2019, where she is currently pursuing the master's degree. Her current research interest includes signal processing.

YUJIE BAI received the bachelor's degree in software engineering from Xinjiang University, China, in 2020, where she is currently pursuing the master's degree in electronic information. Her current research interest includes medical image processing.

XIAOYI LV received the M.S. degree in information and communication engineering from Xinjiang University, China, in 2006, and the Ph.D. degree in electronic and information engineering from Xi'an Jiaotong University, China, in 2010. He is currently a Professor with the School of Software, Xinjiang University. His current work is focused upon bioinformatics and artificial intelligence in medical diagnosis.

PANYUN ZHOU is currently pursuing the degree in software engineering with the College of Software, Xinjiang University.

MINGRUI MA received the master's degree in computer technology from the Dalian University of Technology, in 2012. He has been working with the Information Management Department, The Affiliated Cancer Hospital of Xinjiang Medical University, since 2005. His research interests include computer networks, databases, information security, and artificial intelligence technologies.

● ● ●