Remote Sensing Image Classification Using Deep–Shallow Learning

Peng Dou ⬤, Huanfeng Shen ⬤, *Senior Member, IEEE*, Zhiwei Li ⬤, Xiaobin Guan ⬤, and Wenli Huang, *Member, IEEE*

Abstract—**Recently, classification using multiple classifier system (MCS) has been reported as an effective method to improve remote sensing (RS) image classification. Such systems provide a complementary mechanism to use multiple classifiers, which have shallow architecture to solve the same classification problem; however, the system exhibits shortcomings due to complex ensemble strategy. Deep learning (DL) has been proven to be an advanced method for complex data classification; however, how to use its advantages to overcome the shortcomings of MCS in ensemble strategy for classification accuracy improvement is worthy of study. Thus, with the multiple classifier mechanism and DL architecture, we propose a novel RS image classification framework, namely, deep–shallow learning (DSL), to improve classification accuracy. The DSL framework consists of a shallow learning (SL) layer and a DL layer. The SL layer contains various classifiers with shallow architecture, which can output different classification results for a certain input, whereas the DL layer is formed by DL networks, which can continue learning from the outputs of the SL layer. DSL simulates a human thinking model that continuously learns from the existing learnings to improve learning efficiency. In our experiment, three shallow classification algorithms, i.e., C4.5, k-nearest neighbor, and naive Bayesian, are used to train base classifiers in the SL layer, whereas a deep belief network (DBN) is used to train the DL layer. The experiment results on three different datasets indicate that DSL outperforms other methods in terms of classification accuracy by using backpropagation neural network, bagging, AdaBoost, random forest, multilayer perceptron, and DBN.**

Index Terms—**Deep–shallow learning (DSL), deep learning (DL), ensemble learning (EL), image classification, remote sensing (RS).**

I. INTRODUCTION

REMOTE sensing (RS) technology provides the most intuitive approach to observe land use and land cover (LULC), which are important for studies on climate change, earth system energy balance, and human sustainable development [1]–[5]. Over the past two decades, the use of RS image classification with machine learning to map LULCs with high accuracy and efficiency has gained popularity [6]–[9]. Typical classification algorithms, such as maximum likelihood, decision tree, support vector machine, naive Bayesian (NB), and k-nearest neighbor (KNN), have been widely used to produce LULC data in the early stage and exhibit satisfactory performance [7], [12]–[15]. However, none of these algorithms can produce perfect classification for all LULC categories due to their single-classifier structure; thus, advanced algorithms are urgently needed to meet the requirements of complex RS image classification.

Multiple classifier systems (MCS) [also known as ensemble learning (EL)], which are a powerful method of RS image classification, are based on the theory that multiple weak classifiers can be equivalent to a strong classifier with high generalization and accuracy; thus, it can solve problems, such as small training sample set and local optimum, which weak classifiers cannot address [16]–[20]. Weak classifiers, which are also called base classifiers, are diverse and differ from each other. Their outputs are complementary and can be used to improve classification performance [16], [21]–[22]. Recently, advanced MSCs have been reported to be effective methods to improve RS image classification. For example, Samat *et al.* used ExtraTrees and maximally stable extremal region-guided morphological profile to classify very-high resolution images and achieved good performance for urban mapping [23]. They also introduced CatBoost for RS image classification using diverse features and showed that CatBoost has better capability of reducing the overfitting issue at large number of boosting iteration [24]. Meanwhile, some other MCS algorithms like Light Gradient Boosting Machine, Meta-XGBoost have been proved to be very effective approach to extract information from RS images [25].

In an MCS, a rational ensemble strategy, which combines advantages of different base classifiers, plays an important role [21], [26]–[27]. With such strategy, the differences existing in the base classifiers are fully used to advance the classification performance. To date, these ensemble strategies include the statistical and voting methods. The statistical method builds on prior knowledge to obtain the relations of different base classifiers, whereas the popular method includes Bayesian average, maximum probability, fuzzy theory, and consistency theory [16], [18], [26]. The voting method obtains results via weighting or simple vote the outputs of base classifiers; some popular MCSs, including AdaBoost (AB), rotation forest, random forest (RF) and bagging (BA), are founded on such strategy [29]–[32]. However, collecting enough samples to evaluate the properties of base classifiers and set up a stable ensemble strategy is difficult due to the effect of natural conditions and other factors [19], [33]; as a result, research on MCS must explore the complex relationships among base classifiers. Thus, several advanced

ensemble strategies have been proposed by current research to reflect the relationship of different base classifiers in MCS accurately.

Recently, deep learning (DL) has been considered one of the most effective methods for RS image classification [34]–[38]. It uses brain simulations to establish deep structures with multiple layers to extract high-level features progressively and solves complex classification problems. Compared with classifiers that belong to shallow learning (SL) methods, DL has advantages to learn features automatically and classify them using deep networks to improve accuracy [36],[38]. DL has better generalization capacity compared with SL [38]–[39]. Well-known methods, such as deep belief networks (DBN), recurrent neural networks (RNN), and convolutional neural networks (CNN), have been widely used for LULC mapping [40]–[42]. Various DL methods have different aspects when they solve different classification problems. Some DL methods, such as CNN, are focused on the powerful capability of feature learning [36],[38]. By contrast, some DL methods, such as DBN and RNN, are focused on classification with deep architectures. For example, [40] improved DBN via pretraining and fine-tuning, applied the network to classify hyperspectral RS images and obtained better performance than the original method. Huang *et al.* [41] proposed a semitransfer deep CNN to overcome the shortcomings of traditional CNN and extracted high-accuracy LULU from World View-3 and WorldView-2 images. Lyu *et al.* [42] programmed an RNN model to minimize the seasonal spectral differences of urban areas and obtained the LULC data of four sites located in different regions by using limited training samples.

Continuous learning from existing learnings (CLFEL) is a common thinking model and an effective way to improve learning efficiency. DL is priory for solving complex classification problems; however, most of them take RS images as material, and only a few recognizes the idea of CLFEL [34]–[38]. MCS uses multiple learners, and some researches fused various DL networks to improve classification performance [43], [44]. However, but up to now, research that combines multiple learners with DL methods is rare. Thus, the multiple classifiers' mechanism and powerful cognitive capability of DL must be maximized to achieve a new learning model based on the idea of CLFEL and thus achieve an improved classification performance. From the perspective of MCS, the new model attempts to build a relationship among different base classifiers by using the DL strategy. From the perspective of DL, the outcomes of each base classifier in the MCS are used as input features to train a DL network, which performs deeper learning on the existing learning results. Based on idea of CLFEL, we proposed a new concept called deep–shallow learning (DSL), which inherits the characteristics of MCS and DL. DSL consists of an SL and a DL layer. The SL layer contains various classifiers with shallow architecture, which can output different classification results for a certain input; the DL layer is formed by DL networks, which can continuously learn from the outputs of the SL layer. Thus, the main objectives of this article are as follows: 1) proposing a framework which combines the MCS and DL to improve the RS classification performance, 2) developing an effective feature fusion method to make the outputs of base classifiers adapt to DL layer construction, and 3) investigating the feasibility of CLFEL on RS classification and achieving accurate LULC mapping results.

The remainder of this article is organized as follows. Following the introduction in Section I, Section II describes DSL framework. Section III presents experiment and results. Section IV discusses the diversity of base classifiers in DSL, fusion of classification results with spectral features, classification accuracy improvements, and efficiency. Finally, Section V concludes the article.

II. METHODOLOGY

A. DSL Framework

Fig. 1 illustrates the DSL classification method of this article. The classification model can be divided into two levels. The first level is called SL layer, which consists of various base classifiers produced by different training subsets. The outputs of base classifiers are fused with the spectrum to generate new features, which are considered the output of the SL layer. The second layer is called DL layer, which executes classification using the features provided by the SL layer. The specific process of model training is as follows.

1) Samples are selected from the RS image to generate a training set. Then, 80% of the samples are repeatedly selected from the training set via random replacement sampling to produce a list of subtraining sets. With the subtraining sets, various base classifiers are trained.
2) The base classifiers are used to classify the dataset with 100% samples. Then, the classification results are fused with the image spectrum to produce a new training set in accordance with the method mentioned in Section II-C. In the new training set, samples have the same location with the original training set on the RS image.
3) The new training set is used to produce a DL network to construct the DL layer of DSL. In this article, we use the DBN to build the DSL model.

With the DSL model, a pixel on the RS image is first classified by the base classifiers in the SL layer; then, the classification results are fused with the spectrum to generate new features. These new features are then taken as input of the DL layer, and the pixel is reclassified as the final output. In this manner, when all pixels are processed, accurate LULC maps are produced.

B. SL Layer Construction

SL methods usually have only one hidden layer or no hidden layers; their classifiers have a simple structure and fast running speed and can be trained in a short time [34], [52]–[53]. In this article, SL algorithms with little or no parameters (e.g., C4.5, KNN, and NB) are used to train base classifiers in the SL layer, thereby reducing complexity.

The diversity of base classifiers is the foundation of EL [21], [26]–[27]. The methods of base classifiers are applied equally to produce base classifiers in DSL because for the same object, the differences of the classification results of base classifiers can be used to reflect the properties of each base classifier. BA is a popular EL framework which uses random sampling with replacement to produce different base classifiers. In this article, we refer to the BA method to propose an approach for building the SL layer of DSL. The process of production of base classifiers in the SL layer is shown in Fig. 1. In each epoch, 80%
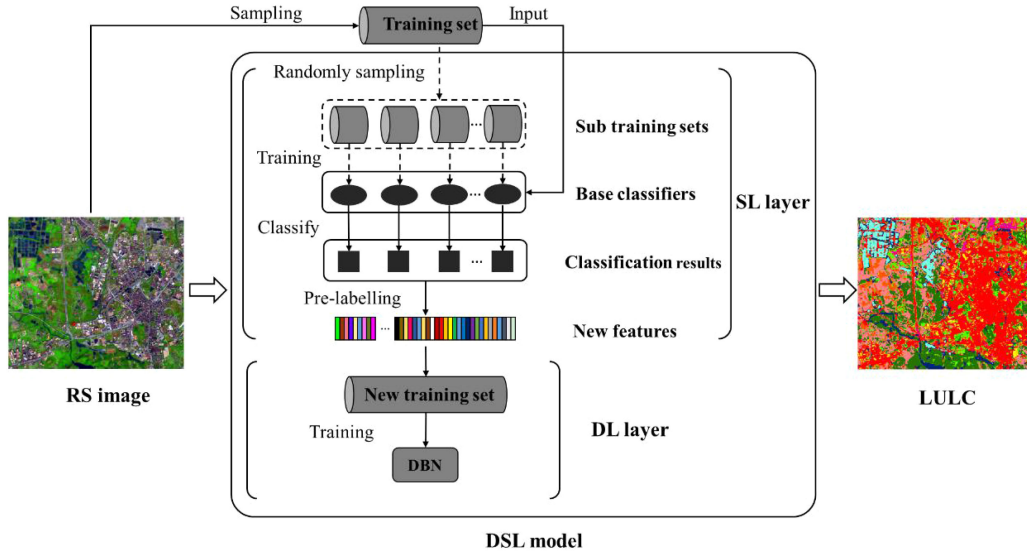
Fig. 1.   DSL classification flowchart.

of the samples are randomly selected from the training set with replacement to generate a subset to train a base classifier. After several epochs, a series of versatile base classifiers is trained for the SL layer. Then, the samples in the training set were classified by the base classifiers, thus generating a new dataset which contains class labels.

C. Fusion of Classification Results and Spectral Features

For an input sample, the results of the SL layer can be recorded as a vector whose entries correspond to the class labels outputted by each base classifier. However, the class labels are discrete codes, which lack the property of continuous values, which may not contribute to high-level classification in the DL layer. Fortunately, the spectrum information of RS images consists of continuous values. Thus, if the class codes and spectral information are fused, then new features that contain predicted class information and spectrum information are generated. Therefore, we propose a prelabeling method to combine class codes and spectrum information and desire useful features from the classification results of SL classifiers.

Let $i = \{12, \ldots, n\}$, $j = \{12, \ldots, l\}$, and $t = \{12, \ldots, c\}$, where n is the number of classifiers in the SL layer, l is the number of bands of the RS image, and c is the number of classes. For each pixel of RS image, the spectral features can be described by $D = [d_1, d_2, \ldots, d_l]$. Taking D as the input of the classifier H_i, the output of H_i can be considered a class code r_i, where $r = \{12, \ldots, c\}$. To recognize the classification results from different classifiers, a unique random number $y_i \in (01)$ is generated for classifier H_i, and the final output of classifier H_i can be regarded as $r_i y_i$. Then, if D is taken as the input of classifiers set $H = \{H_1, H_2, \ldots, H_n\}$, the outputs of H can be expressed as $R = [\, y_1 r_1, y_2 r_2, \ldots, y_n r_n]^{\mathrm{T}}$. With the results of H, the pixel described by D can be prelabeled by using (1):

$$F = RD = \begin{bmatrix} r_1 y_1 d_1 & r_1 y_1 d_2 & \cdots & r_1 y_1 d_l \\ r_2 y_2 d_1 & r_2 y_2 d_2 & \cdots & r_2 y_2 d_l \\ \vdots & \vdots & \ddots & \vdots \\ r_n y_n d_1 & r_n y_n d_2 & \cdots & r_n y_n d_l \end{bmatrix} \quad (1)$$
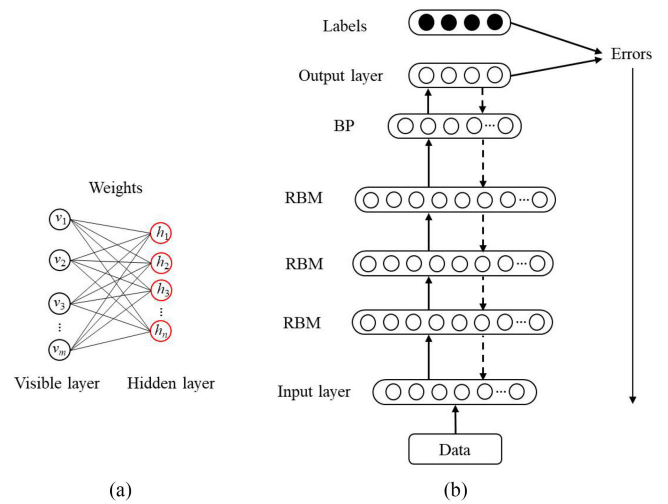

Fig. 2.   Training process of DBN.

Then, F is rearranged by row to obtain a vector F' with $n \times 1$ dimensions, representing the new features of the current pixel.

After prelabeling the entire RS image, a new image with $n \times l$ channels is obtained. The new image contains spectral and SL classification information and has additional features for classification in the DL layer.

D. DL Layer Construction

In DSL, DBN is used to continue learning from the outputs of the SL layer and thus obtain more accurate classification results. As reported in many studies, DBN has been considered one of the most popular DL networks in the field of RS image classification. As shown in Fig. 2(a), a DBN model can be viewed as a composition of several layers of restricted Boltzmann machines (RBMs) and a layer of back propagation (BP) neural network. RBM is an unsupervised network which consists of visible and hidden layers. The hidden layer serves as the visible layer for the next, and a pair of the units from either of the two layers has

TABLE I
INFORMATION OF DATASETS

Dataset	Image	Bands	Area（km²）	Centre coordinates	Acquisition date
1	Landsat OLI	1,2,3,4,5,6,7,8	29.670	22°9' N, 114°16' E	2013/01/13
2	Spot-5	2,3,5	4.388	39°7' N, 117°3' E	2011/06/07
3	GaoFen-2	1,2,3,4	4.117	28°46' N, 120°44' E	2013/07/10

a symmetric connection between them; units within a layer have no connections. With the RBM, the probability distributions over its set of inputs can be learned. On the basis of the staking layers of RBMs, the training process of DBN can be divided into the pretraining and fine-tuning stages.

1) Pretraining: RBMs are trained from the first RBM, which consists of the input layer and the first hidden layer. The output of each FRM is treated as the visible layer of the next RBM. The layer-by-layer learning process can be repeated until the last hidden layer.

2) Fine-tuning: After pretraining, the supervised learning and BP neural network are used to propagate the error of the outputs of the last RBM and the labels layer by layer and to adjust the weight and bias with several interactions until the globally optimal DBN is obtained.

As the unsupervised learning process in the pretraining stage can be considered the initialization of the BP network instead of random initialization, the backpropagation in the DBN does not fall into the local optimal solution easily. Thus, in this article, we select DBN to construct the DL layer in the DSL framework.

Fig. 3. Remote sensing images.

III. EXPERIMENT AND RESULTS

A. Experiment Dataset

In our experiment, three cloud-free RS images of Landsat OLI, Spot-5, and GaoFen-2 are used to produce dataset-1, -2, and -3, respectively. The information of the RS images is listed in Table I. For dataset-1, the Gram–Schmidt pan sharpening method is used to enhance the spatial resolution of bands 1–7 (30-m spatial resolution) using panchromatic band 8 (15-m spatial resolution). By staking the processed bands 1–7 and band 8, a new multispectral band image with 15 m is obtained. For dataset-2, the same method is used to enhance the multispectral bands (10-m spatial resolution) using the panchromatic band (5-m spatial resolution). The multispectral bands of GaoFen-2 (4-m spatial resolution) are used generate dataset-3. The study plots of the three datasets are illustrated in Fig. 3.

By analyzing the landcover status of the study areas, we establish classification schemes for datasets 1–3. Dataset-1 includes nine classes: grassland (GR), low-reflectivity buildings (LRB), high-reflectivity buildings (HRB), river (RI), dike-pond (DP), bare land (BL), cultivated land (CL), noncultivated land (NCL), and forest. Dataset-2 includes six classes: artificial land cover (ALC), FO, water body (WB), CL, GR, and BL. Dataset-3 includes nine classes: arbor (AR), shrubs (SH), CL, HRB, LRB, WB, hardened ground (HG), BL, and NCL. To facilitate data processing, the class labels of all datasets are digitized by successive integers starting from 1; these integers represent
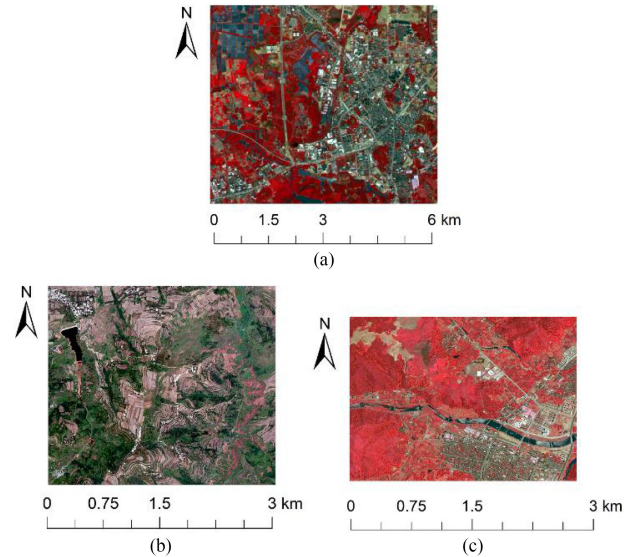
different classes. The specific digitization code of each class label is shown in Table II.

High-resolution images from Google Earth are used as a reference to select the training samples and test samples of each dataset. All samples are point-based and the training samples are typical pure pixels. The test samples are randomly selected pixels, and they are then assigned to the class labels using the reference image. The number of samples of different datasets is shown in Table III.

B. Parameter Setting

In our experiment, the number of classifiers in the SL layer was set as 50. The parameters of C4.5, NB, and KNN, which were used for training base classifiers in the SL layer, and DBN, which was used for training deep networks in the DL layer, are described in Table IV. To evaluate the performance of DSL the framework, several excellent classification algorithms, including back propagation neural network (BPNN), RF, AB, BA, and multilayer perceptron (MLP), were used for comparison; their parameters are also described in Table IV. For all the algorithms, the parameters are optimized with practical experience, thereby stabilizing the performance of the algorithms in the experiment.

C. Classification and Accuracy Analysis

Pixel-based classifications were executed on the three datasets with DSL and other algorithms. Within the classification

TABLE II
DIGITIZATION CODES OF CLASSIFICATION SCHEMES

Dataset	Classification schemes
1	1(GR), 2(LRB), 3(HRB), 4(RI), 5(DP), 6(BL), 7(CL), 8(NCL) and 9(FO)
2	1(ALC), 2(FO), 3(WB), 4(CL), 5(GR) and 6(BL)
3	1(AR), 2(SH), 3(CL), 4(HRB), 5(LRB), 6(WB), 7(HG), 8(BL) and 9(NCL)

TABLE III
NUMBER OF TRAINING AND TEST SAMPLES

Dataset	Class label	Number of training samples	Number of test samples
Dataset-1	GR	116	367
	LRB	138	416
	HRB	118	295
	RI	116	357
	DP	123	376
	BL	124	251
	CL	130	263
	NCL	114	306
	FO	115	377
Dataset-2	ALC	232	701
	FO	217	681
	WB	105	612
	CL	195	712
	GR	196	773
	BL	198	600
Dataset-3	AR	100	500
	SH	100	500
	CL	100	500
	HRB	100	500
	LRB	100	500
	WB	100	500
	HG	100	500
	BL	100	500
	NCL	100	500

algorithms, BA and AB are two EL frameworks which can be worked on different supervised classification algorithms. Here, we used C4.5, KNN, and NB to train the base classifiers of BA, AB, and DSL. The classification accuracy of all methods is summarized in Table V.

When single classifiers (SC) are used, the accuracy of C4.5, KNN, and NB is low. The OA of dataset-1, -2, and -3 are 75.69%–80.47%, 77.57%–82.18%, and 63.78%–72.61%, respectively. However, when it comes to DSL which uses C4.5, KNN, and NB to produce the base classifiers, the accuracy for the three datasets reaches 86.91%–87.7%, 88.15%–88.97%, and 80.53%–84.18%, respectively. DSL exhibits higher accuracy compared with its base classifiers and the SC of DBN, which have accuracies of 85.54%, 85.28%, and 75.89% for dataset-1, -2, and 3, respectively. Moreover, the contrastive classification methods have higher accuracy than that of C4.5, KNN, and NB but lower accuracy than that of DSL. This feature indicates that DSL performs better than the existing methods in the experiment in terms of improving RS image classification.

IV. DISCUSSION

A. Diversity of Base Classifiers in DSL

For the same classification problem, different classifiers have different classification results. Some researchers have reported that even though no remarkable difference exists among different classifiers at the overall level, such dissimilarities can be found at the per-class level [16]–[19], [21]–[22]. Such situation is suitable for the base classifiers of DLS. To evaluate these

TABLE IV
DESCRIPTION OF THE MAJOR PARAMETERS OF DIFFERENT CLASSIFIERS

Classification algorithm	Parameter description
C4.5	Minimum number of samples required to split an internal node: 2; Minimum number of samples required to be at a leaf node: 1.
KNN	Number of neighbours: 5; Power parameter for the Minkowski metric: 2.
DBN	Number of epochs for the pretraining and fine-tuning are respectively set as 200 and 500. Three layers of RBMs with 50 hid units are constructed, and the hid units of the fine-tuning is 200. The learning ratios of pretraining and fine-tuning are set as 0.001, and the batch size is set as 30.
RF	Number of trees in the forest: 20; Minimum number of samples required to split an internal node: 2; Minimum number of samples required to be at a leaf node:1.
MLP	Number of epochs: 1000; List of hid unit count: [100,100,100]; Learning ratio: 0.001; Batch size: 30;
BPNN	Number of epochs: 1000 Hid unit count: 100 Learning ratio: 0.001 Batch size: 30

TABLE V
CLASSIFICATION ACCURACY OF DIFFERENT CLASSIFICATION METHODS ("–" MEANS NO VALUE EXISTS IN THE CELLS)

| Dataset	Algorithm	Single classifier		DSL		BA		AB	
		OA (%)	Kappa	OA (%)	Kappa	OA (%)	Kappa	OA (%)	Kappa
Dataset-1	NB	75.69	0.616	**86.91**	**0.767**	84.01	0.742	82.66	0.731
	KNN	78.04	0.644	**87.21**	**0.766**	83.84	0.734	83.68	0.734
	C4.5	80.47	0.680	**87.70**	**0.779**	83.64	0.733	84.22	0.738
	DBN	85.54	0.722	–	–	–	–	–	–
	MLP	82.38	0.701	–	–	–	–	–	–
	RF	81.69	0.698	–	–	–	–	–	–
	BPNN	83.78	0.733	–	–	–	–	–	–
Dataset-2	NB	82.18	0.693	**88.97**	**0.838**	83.4	0.813	83.65	0.739
	KNN	81.96	0.687	**88.59**	**0.849**	80.22	0.778	81.26	0.782
	C4.5	77.57	0.632	**88.15**	**0.828**	82.28	0.726	83.21	0.735
	DBN	85.28	0.782	–	–	–	–	–	–
	MLP	85.99	0.781	–	–	–	–	–	–
	RF	83.58	0.715	–	–	–	–	–	–
	BPNN	83.37	0.763	–	–	–	–	–	–
Dataset-3	NB	72.61	0.692	**84.18**	**0.822**	81.40	0.813	82.71	0.806
	KNN	69.06	0.652	**80.53**	**0.789**	76.91	0.774	78.51	0.758
	C4.5	63.78	0.592	**80.87**	**0.792**	75.96	0.775	76.09	0.731
	DBN	75.89	0.725	–	–	–	–	–	–
	MLP	76.20	0.729	–	–	–	–	–	–
	RF	76.93	0.628	–	–	–	–	–	–
	BPNN	79.29	0.730	–	–	–	–	–	–
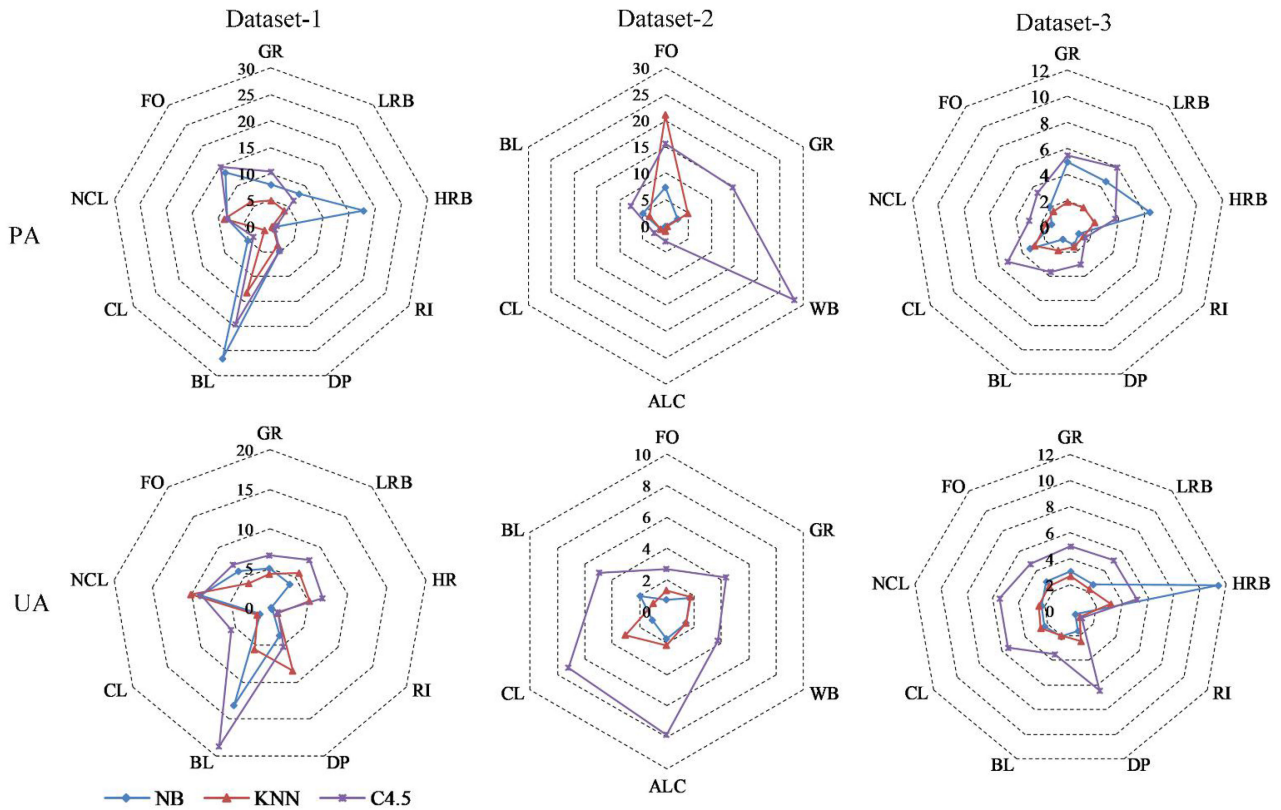
Fig. 4.   SD of classification accuracy at the per-class level in the SL layer.

TABLE VI
SD OF THE CLASSIFICATION ACCURACY

Dataset	Classifier	SD	
		OA	Kappa
	NB	3.409	0.042
Dataset-1	KNN	1.140	0.017
	C4.5	1.696	0.026
	NB	0.971	0.013
Dataset-2	KNN	1.493	0.021
	C4.5	2.611	0.032
	NB	1.760	0.020
Dataset-3	KNN	0.699	0.008
	C4.5	1.529	0.017

TABLE VII
PERCENTAGE OF THE INERT PIXELS OF THE ENTIRE IMAGE OBTAINED BY
DIFFERENT SL CLASSIFIERS (%)

	NB	KNN	C4.5
Dataset 1	61.95	75.67	25.00
Dataset 2	5.74	35.28	10.65
Dataset 3	2.95	5.99	3.42

differences, standard deviations (SDs) of the OA are used to measure the dispersion of the base classifiers' results in the DLS (Table VI). The higher the SD, the more differences among these classifiers would have. Furthermore, the SDs of the OA for the base classifiers are greater than 1, indicating differences and diversity among the base classifiers in the DSL framework. Fig. 4 illustrates the SD of classification accuracy at the per-class level. Under the same classification algorithm, the SD of user accuracy (UA) and producer accuracy (PA) is substantially larger than the SD of OA, indicating that the differences among base classifiers at the per-class level are larger than that of the overall level.

With the base classifiers, 50 LULC maps are generated and digitized with the schedule listed in Table II. Then, the average value for each pixel is calculated to obtain an average image (AI). The grey histograms of these AIs are shown in Fig. 5. For each pixel of the AI, if the value is closer to the class label value, then the outputs of the base classifiers become more consistent. For convenience, pixels whose all base classifiers have the same classification results are referred to as inert pixels, and the other pixels are called noninert pixels. The percentages of the inert pixels of the entire image obtained by different base classifiers are listed in Table VII. When the base classifiers in the DSL are used to classify the RS image, most pixels are noninert pixels, and the outputs of these classifiers vary. To some degree, the properties of different objects under each classifier can be exhibited by the classification results. That is, when the outputs of all base classifiers in the DSL are distributed into feature space, the combination of different base classifiers' outputs can be used for object identification.
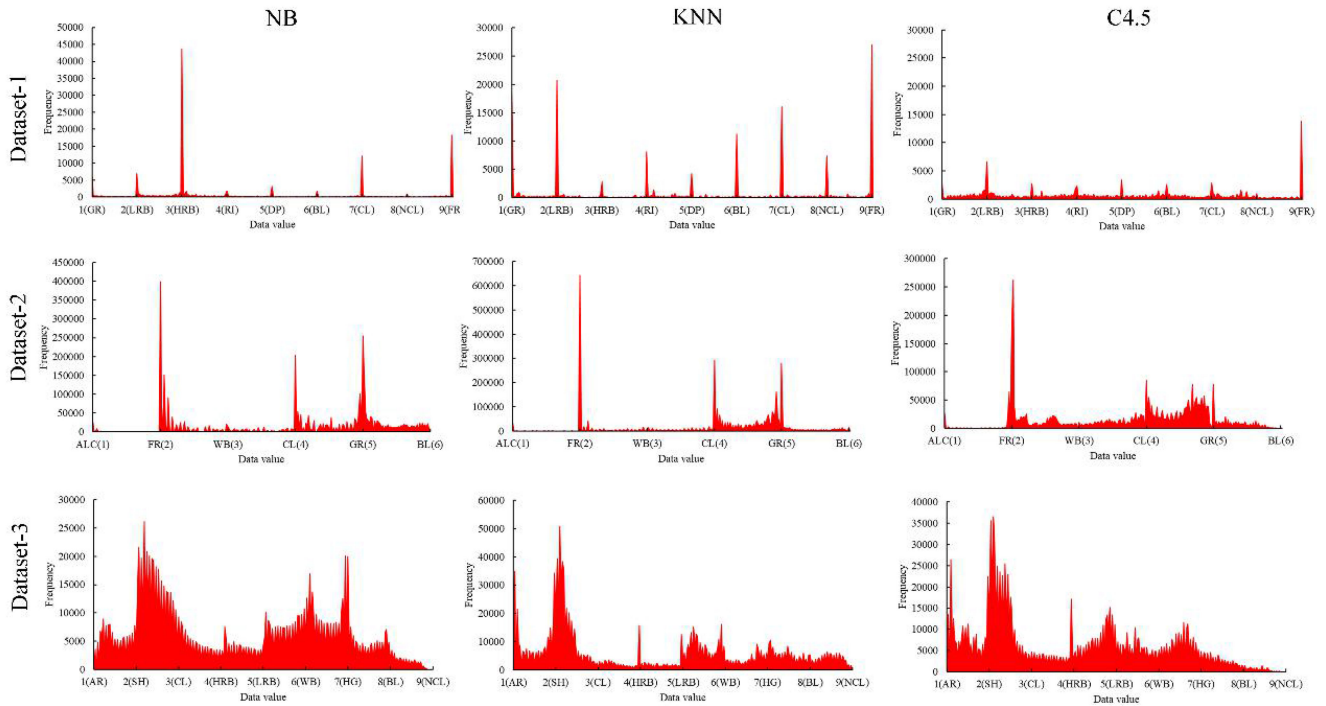
Fig. 5.   Histograms of AIs with the use of different SL algorithms.

B. Fusion of Classification Results With Spectral Features

In the DSL, the proposed prelabeling method fuses the outputs of the base classifiers with the spectral information. Theoretically, the new features contain abundant information for distinguishing different objects. To explore the characteristics of these features intuitively, 300 samples for each class are randomly selected, and these samples are arranged from smallest to largest according to the class label values. In this manner, a scatter plot graph between features and classes is generated, as shown in Fig. 6. Different classes show different stripes on the graph with particular layering. This feature makes objects on the RS image more easily distinguished. For example, in dataset-1, DP and RI, which are difficult to distinguish, show some similarities in the RS image; however, with the fused features, the differences between them are enhanced, thus reflecting a fairly optimistic separability on the graph. The properties exhibited on the graphs show that the prelabeling method provides additional optimal features, and this feature can be viewed as an important basis for the classification improvement in DSL.

C. DSL Classification Performance

1) Analysis of the Classification Performance of DSL: DSL exhibits high classification accuracy because of the improvements at the per-class level. Fig. 7 shows the PAs and UAs of DSL and DBN whose accuracy can be viewed as the representation of comparative methods. Except for a few classes, DSL has improved classification accuracy at the per-class level, indicating that most individual classes predicted by DSL are closer to the ground truth.

DSL uses an MCS to solve classification problems; thus, the types and counts of base classifiers play an important role in

classification improvements. Fig. 8 illustrates the classification accuracy improvements with base classifiers increasing in the DSLs. Some DSLs, such as C4.5-based DSL that worked on dataset-1 and -2 and NB-based DSL that worked on dataset-3, exhibit a continuous increase in accuracy with increase in base classifiers. For some other DSLs, significant improvements in classification accuracy are observed in the early stage, but such improvements are saturated in the later stage, such as KNN-based DSL worked on dataset 1 and KNN-, and NB-based DSL worked on dataset-2. This phenomenon suggests that with different algorithms to produce base classifiers in DSL, classification improvement becomes different. Considering the percentage of inert pixels in RS image classification (Table VII and Fig. 5), we find that for dataset-1, C4.5 produces the lowest inert pixels percentage (25%) in comparison with KNN and NB; thus, C4.5-based DSL has the largest accuracy improvement. For dataset-3, the percentage of inert pixels produced by NB is less than that of C4.5 and KNN, whereas the NB-based DSL exhibits the largest classification improvement. That is, if the base classifiers are more diverse, then these classifiers can facilitate classification accuracy improvement with DSL.

Many aspects can be used to explain why DSL has better classification performance than SL, DL, and other comparative methods. First, DSL uses a series of features extracted from the classification results of base classifiers with SL methods. These features sign class information to the original RS data and increase the recognizability of different classes (Fig. 6), resulting in an easy classification for the DL layer. Second, it is based on the idea of CLFEL, learning from the results of SL classifiers to obtain accurate classification with the DL method. Furthermore, different base classifiers perform differently, and their classification results exhibit certain complementary
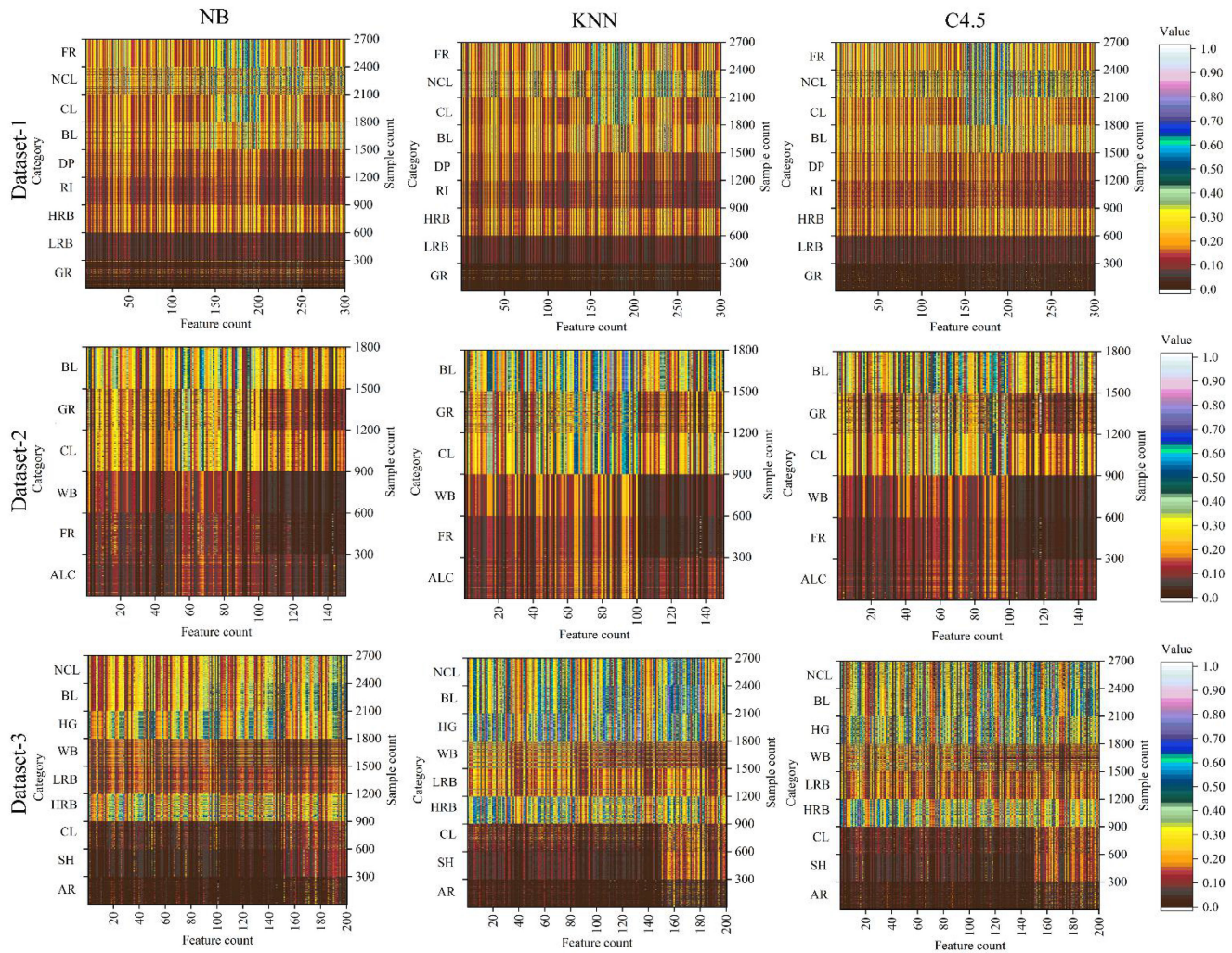
Fig. 6.   Scatter plot graphs between features and classes of using different SL classifiers.

relationships. These complements can be combined by DL in the DSL. In addition, DSL integrates the SL and DL to minimize the problem of insufficient classification by using SC, and it improves classification accuracy effectively.

2) Analysis of Time Consumption: DSL is a time-consuming algorithm due to its MCS. Table VIII lists the training time for different classifiers. The algorithm selection in the SL layer affects the time consumption of DSL. For example, the KNN method needs more time for training compared with C4.5 and NB, making KNN-based DSL works on different datasets consume more time compared with C4.5 and NB-based DSL. Moreover, the classifiers in the SL layer in DSL consume less time compared with the DBN in the DL layer because SL methods, such as C4.5, KNN, and NB, have a simpler structure compared with the DL method, DBN. In DSL, the number of base classifiers increases time consumption, but the experiments suggest that the DL layer consumes more time compared with the SL layer in DSL training. Moreover, with 50 base classifiers, the SL layer costs less than 10% of DSL training time, whereas the DL layer dominates the most part of DSL time consumption. This result shows that compared with the DL method of DBN,

DSL is more time consuming, but the increased time consumption is basically within an acceptable range.

The analyses above indicate that DSL inherits time-consuming characteristics from DL and is an algorithm that improves accuracy by consuming time. In addition to the above reasons, many other factors affect DSL training time; these factors include optimal parameters and DL network settings. In our experiment, the classifiers in the SL layer are generated in a serial manner, thus increasing time consumption.

3) LULC Mapping: With the different classification methods, the LULC of different datasets is mapped. The local details are shown in Figs. 9 –11. For dataset-1 (Fig. 9), DP and RI have similar characteristics on the RS image, making them difficult for classification. With the SC of NB, KNN, and C4.5, most peripheries of the DPs are misclassified as RIs, and some objects mixed in the DPs are incorrectly classified as CL. With RF, MLP, BPNN, BA, BA, and DBN, classification seems to improve; however, some errors still exist in the LULC maps. With the DSL, the outline of DP is clearer than other methods with fewer wrongly classified peripheries. For dataset-2 (Fig. 10), with the SC of NB, KNN, and C4.5 and other comparative methods, the
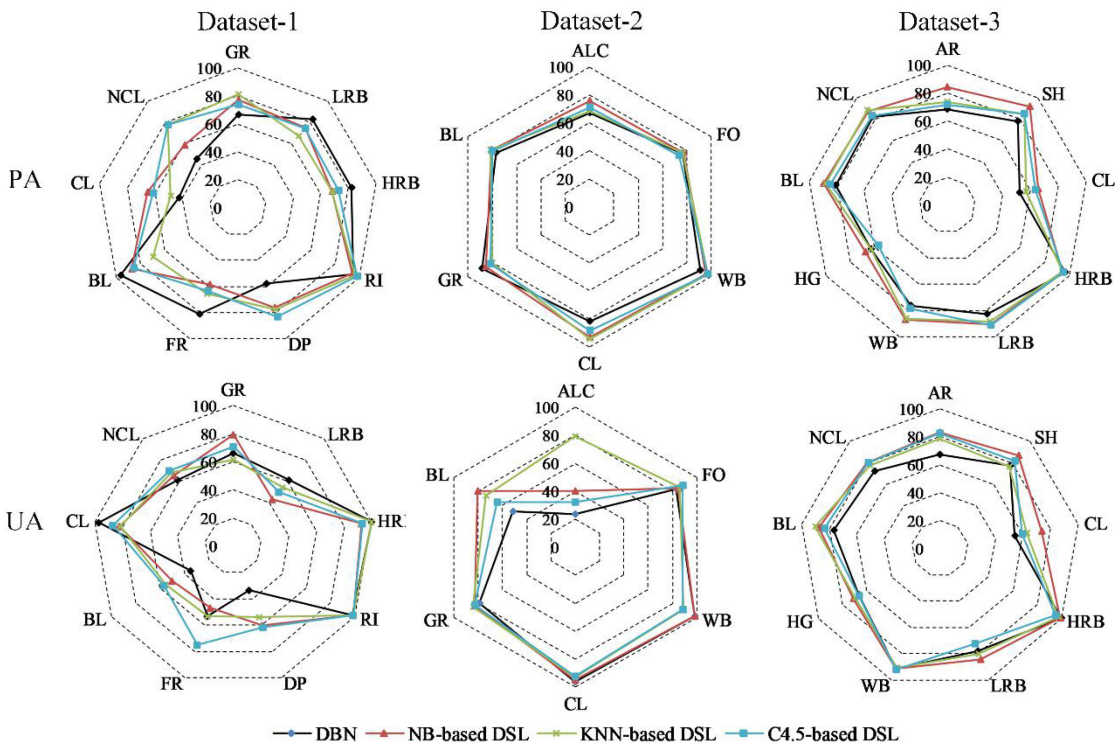
Fig. 7. Classification improvement at the per-class level with the use of DSL.
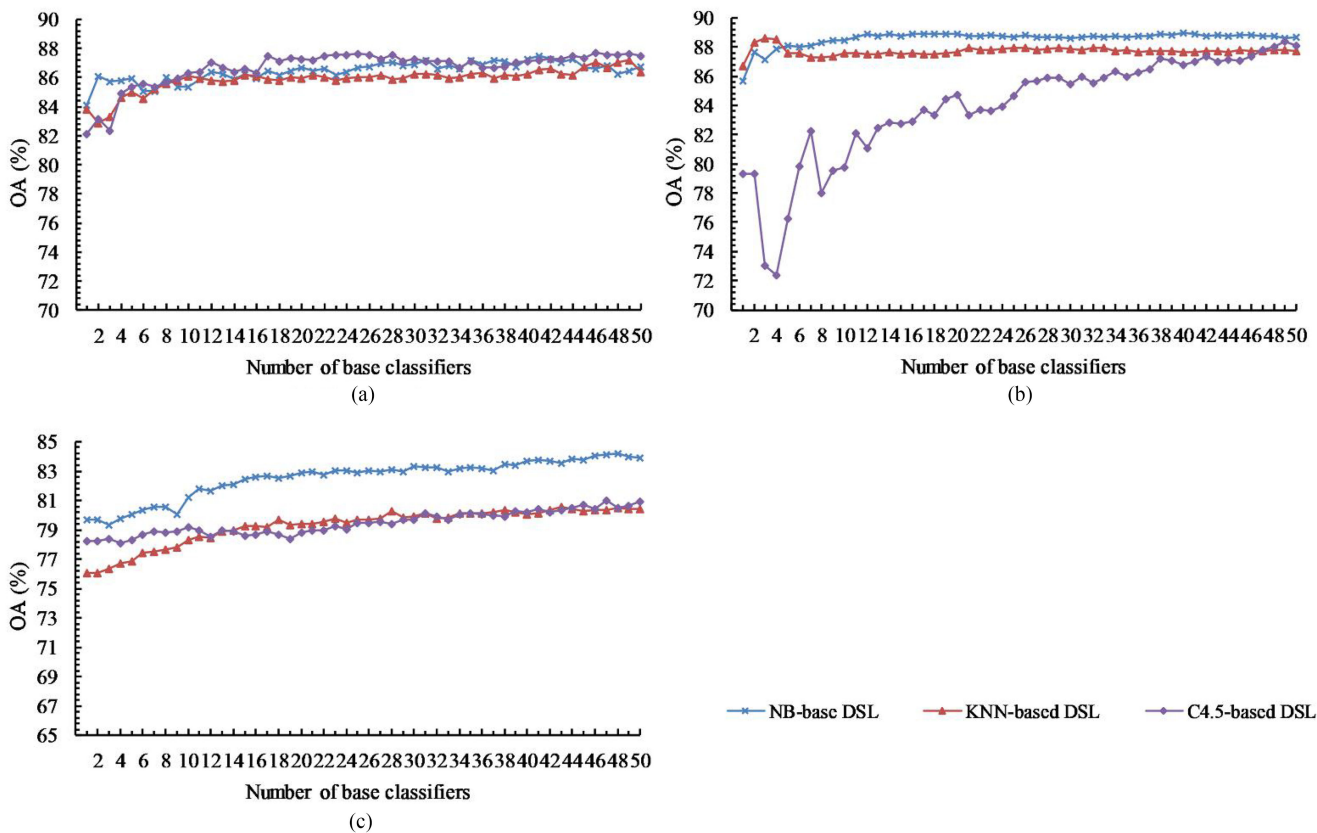

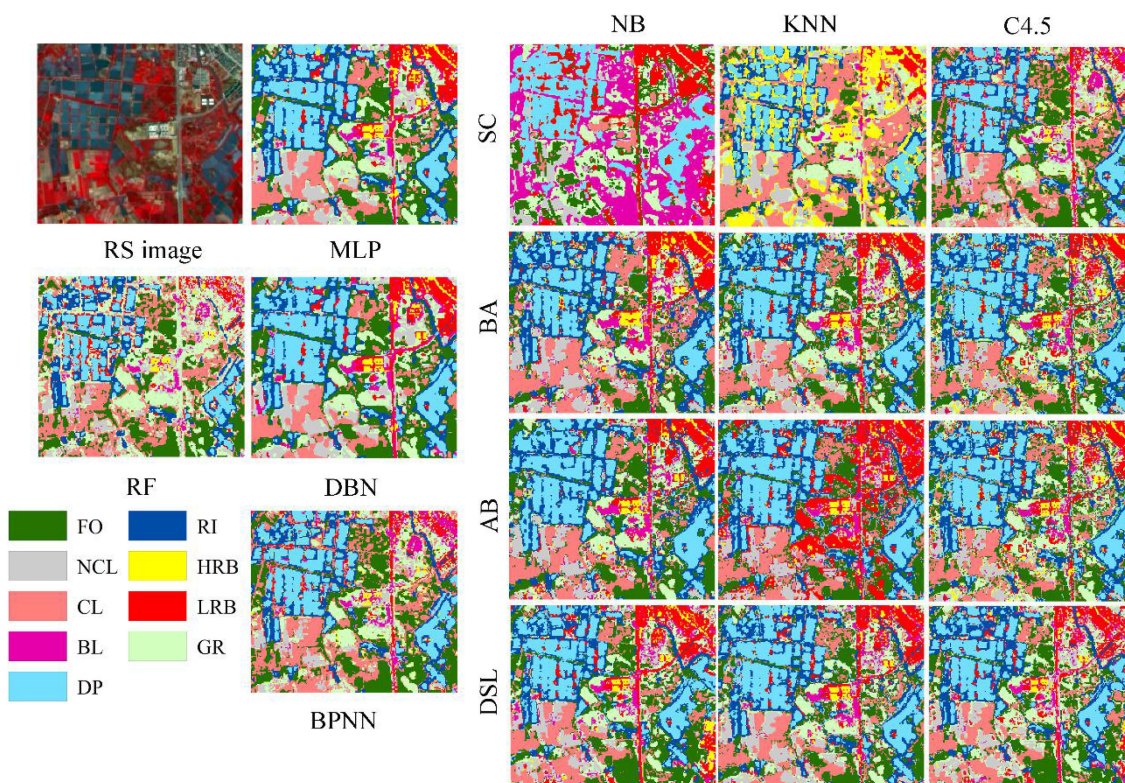Fig. 8. OA improvement via DBN-based DSL and MLP-based DSL.

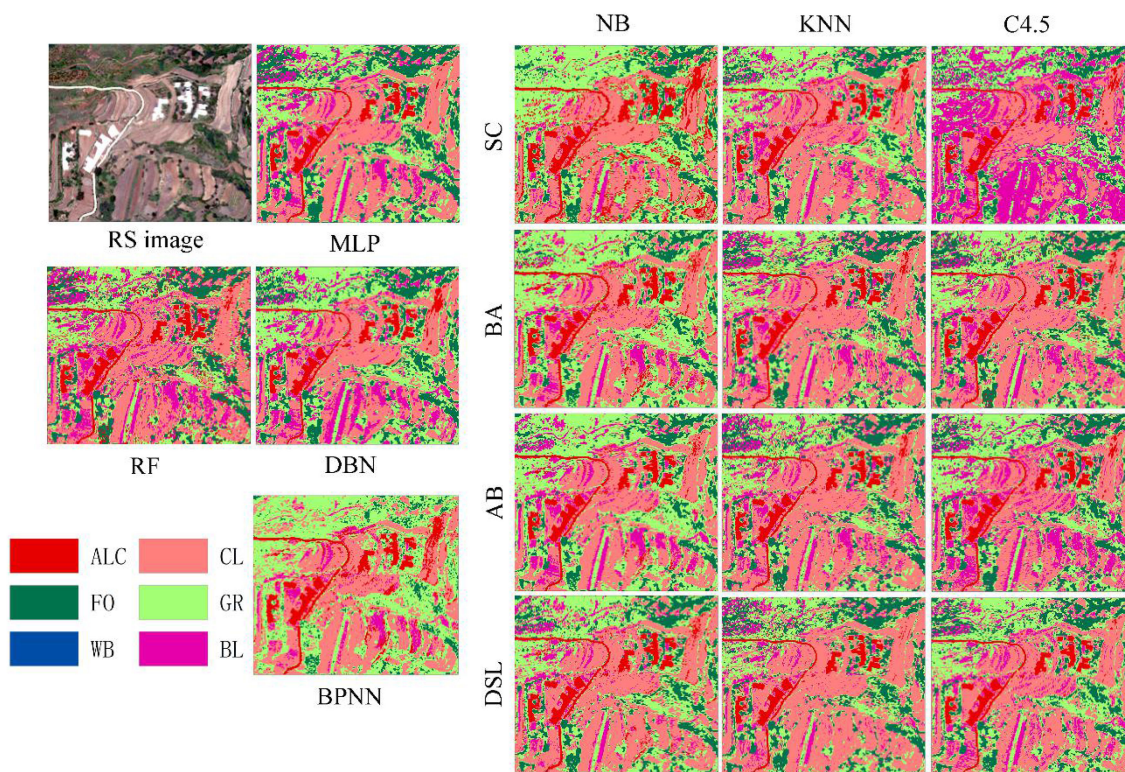Fig. 9. Results of classification using different methods for dataset-1.

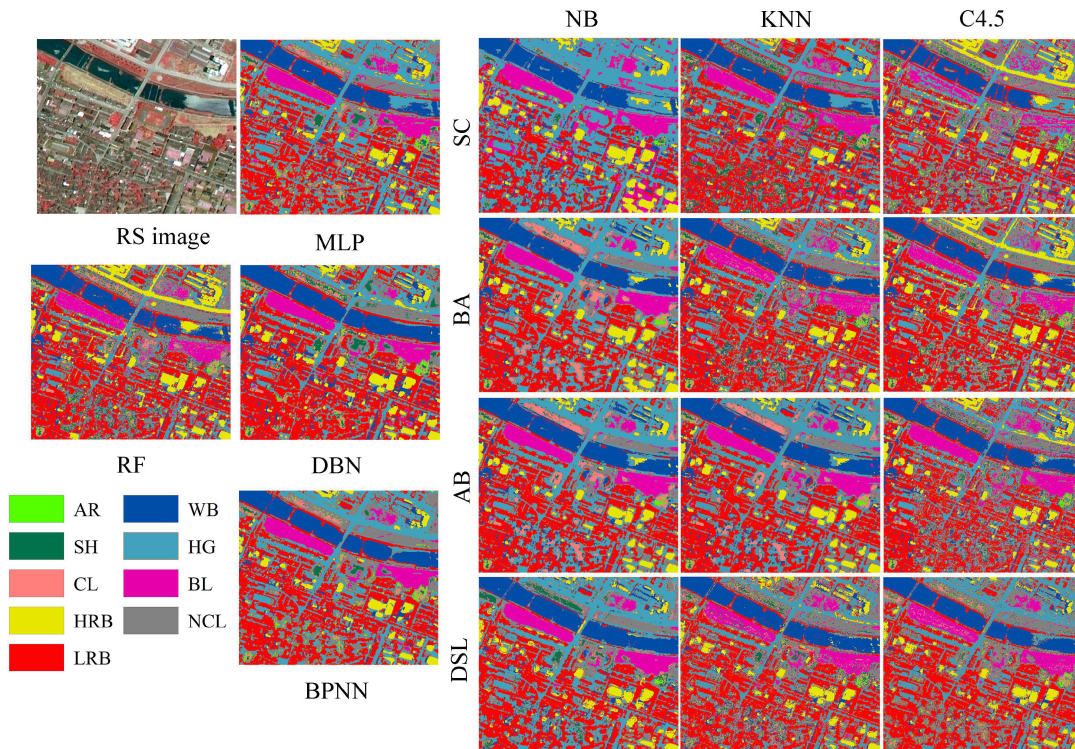Fig. 10. Results of classification using different methods for dataset-2.

Fig. 11.   Results of classification using different methods for dataset-3.

TABLE VIII
TIME CONSUMPTION OF SINGLE CLASSIFIERS AND DSL FOR TRAINING
(SECOND) ("–" MEANS NO VALUE EXISTS IN THE CELLS)

Dataset	Algorithm	SC	DSL
Dataset-1	NB	0.080	46.433
	KNN	0.270	60.815
	C4.5	0.031	57.478
	DBN	38.086	–
Dataset-2	NB	0.036	92.184
	KNN	0.019	94.718
	C4.5	0.014	88.874
	DBN	58.005	–
Dataset-3	NB	0.028	53.181
	KNN	0.023	55.483
	C4.5	0.018	56.420
	DBN	46.644	–

CLs which exhibit bright areas in the image are misclassified as ALCs and BLs at different levels. However, with DSL, most CLs are correctly classified, and DSL has an improved performance to recognize CL, ALC, and BL. In dataset-3 (Fig. 11), the waves in the RI show a similar brightness to HRB or HR, and the shadows casted by the buildings have similar features with the WB. Neither with the SC of NB, KNN, and C4.5 nor with other comparative methods, the waves and shadows disturb the classification results. With DSL, the interference seems minimized, most waves are classified as WB, and the probability that shadows are misclassified as WB is decreased.

The LULC mapping cases of dataset-1, -2, and -3 indicate that DSL has strong power to recognize different classes that SLs, and MCSs that mentioned in this article and the deep architectures difficult to classify. This result is obtained because DSL transforms the varied classification results of base classifiers in the SL layer into series of new features. These features are complementary and can effectively compensate for the defects that some objects cannot be well distinguished from other objects when the spectral information of the original image is insufficient. In addition, DSL realizes some functions of MCS, which can integrate the classification results of different classifiers in the SL layer by using DL. In addition, it obtains more accurate classification results compared with the base classifiers. Furthermore, the essence of DSL is to perform additional DL from the learning results of the SL layer. Compared with the mode which only has one layer of learning mechanism, the multiple learning layer mode, which integrates SL and DL, is another major reason why DSL outperforms other classification methods in LULC mapping.

V. CONCLUSION

On the basis of the idea of CLFEL, the DSL framework is proposed for RS image classification. The tests on three different RS datasets indicate that the DSL effectively maximizes MCS and DL, resulting in higher classification accuracy. Based on the results of the experiments, this article draws the following conclusions:

1) For the same classification problem, the accuracy of different classifiers in the SL layer may not be much different at the overall level, but huge differences are found at

the per-class level, resulting in diverse classifiers whose outputs contain rich information for recognizing different objects. Based on this theory, the classification results and spectral information are fused as new features by using a prelabeling method. This fused feature enhances the characteristics of different objects and exhibits optimistic separability, allowing the objects to be easily classified in the next layer.

2) DSL outperforms other classification methods because of the following reasons: By simulating the cognitive process of CLFEL, DSL has more approaches to improve classification accuracy and overcomes the insufficient classification accuracy of using SC. The new method effectively combines the diversity property of multiple classifiers in the SL layer and the superior classification power of DBN in the DL layer. The multiple learning layer mechanism improves DSL in terms of LULC mapping.

In summary, DSL is a new classification mechanism with shallow and DL layers. This mechanism can be used to improve RS classification remarkably. The framework is flexible for SL classifiers and DL algorithms choosing; thus, it has potential for improvement to generate more accurate LULC maps from RS images. However, DSL is as time consuming as DL; if the framework is designed in a parallel manner, then its operational efficiency would be greatly improved.

ACKNOWLEDGMENT

REFERENCES

[1] M. D. Behera et al., "Remote sensing based deforestation analysis in mahanadi and Brahmaputra river basin in India since 1985," J. Environ. Manage., vol. 206, pp. 1192–1203, Jan. 2018.

[2] H. Tao et al., "Impacts of land use and land cover change on regional meteorology and air quality over the Beijing-Tianjin-Hebei region, China," Atmos. Environ., vol. 189, pp. 9–21, Sep. 2018.

[3] X. Liu et al., "High-resolution multi-temporal mapping of global urban land using landsat images based on the Google Earth Engine platform," Remote Sens. Environ., vol. 209, pp. 227–239, May 2018.

[4] P. Gong et al., "A new research paradigm for global land cover mapping," Ann. GIS, vol. 22, no. 2, pp. 87–102, Mar. 2016.

[5] J. Chen and J. Chen, "GlobeLand30: Operational global land cover mapping and big-data analysis," Sci. China Earth Sci., vol. 61, no. 10, pp. 1533–1534, Sep. 2018.

[6] P. Dou and Y. Chen, "Dynamic monitoring of land-use/land-cover change and urban expansion in Shenzhen using landsat imagery from 1988 to 2015," Int. J. Remote Sens., vol. 38, no. 19, pp. 5388–5407, Jun. 2017.

[7] Q. Shi et al., "Domain adaption for fine-grained urban village extraction from satellite images," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 8, pp. 1430–1434, Aug. 2020.

[8] S. S. Rwanga and J. M. Ndambuki, "Accuracy assessment of land use/land cover classification using remote sensing and GIS," Int. J. Geosci., vol. 08, no. 04, pp. 611–622, 2017.

[9] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," Int. J. Remote Sens., vol. 39, no. 9, pp. 2784–2817, Feb. 2018..

[10] S. Niazmardi, B. Demir, L. Bruzzone, A. Safari, and S. Homayouni, "Multiple kernel learning for remote sensing image classification," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 3, pp. 1425–1443, Mar. 2018.

[11] A. Elshamli, G. W. Taylor, A. Berg, and S. Areibi, "Domain adaptation using representation learning for the classification of remote sensing images," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 10, no. 9, pp. 4198–4209, Sep. 2017.

[12] R. K. Tiwari, M. K. Arora, and R. P. Gupta, "WITHDRAWN: Comparison of maximum likelihood and knowledge-based classifications of debris cover of glaciers using aster optical-thermal imagery," Remote Sens. Environ., Nov. 2014.

[13] B. Mack and B. Waske, "In-depth comparisons of maxent, biased SVM and one-class SVM for one-class classification of remote sensing data," Remote Sens. Lett., vol. 8, no. 3, pp. 290–299, Dec. 2016.

[14] G. Alimjan, T. Sun, H. Jumahun, Y. Guan, W. Zhou, and H. Sun, "A hybrid classification approach based on support vector machine and K-Nearest neighbor for remote sensing data," Int. J. Pattern Recognit. Artif. Intell., vol. 31, no. 10, Mar. 2017, Art. no. 1750034.

[15] P. Du et al., "Advances of four machine learning methods for spatial data handling: A review," J. Geovis. Spatial Anal., vol. 4, no. 1, May 2020, Art. no. 13.

[16] P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu, "Multiple classifier system for remote sensing image classification: A review," Sensors, vol. 12, no. 4, pp. 4764–4792, Apr. 2012.

[17] K. Tan, X. Jin, A. Plaza, X. Wang, L. Xiao, and P. Du, "Automatic change detection in high-resolution remote sensing images by using a multiple classifier system and spectral–spatial features," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 9, no. 8, pp. 3439–3451, Aug. 2016.

[18] Y. Chen, P. Dou, and X. Yang, "Improving land use/cover classification with a multiple classifier system using adaboost integration technique," Remote Sens., vol. 9, no. 10, Oct. 2017, Art. no. 1055.

[19] E. Isaac, K. S. Easwarakumar, and J. Isaac, "Urban landcover classification from multispectral image data using optimized adaboosted random forests," Remote Sens. Lett., vol. 8, no. 4, pp. 350–359, Jan. 2017.

[20] X. Du and A. Zare, "Multiple instance choquet integral classifier fusion and regression for remote sensing applications," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 5, pp. 2741–2753, May 2019.

[21] A. Mellor and S. Boukir, "Exploring diversity in ensemble classification: Applications in large area land cover mapping," ISPRS J. Photogramm. Remote Sens., vol. 129, pp. 151–161, Jul. 2017.

[22] P. Dou, Y. Chen, and H. Yue, "Remote-sensing imagery classification using multiple classification algorithm-based adaboost," Int. J. Remote Sens., vol. 39, no. 3, pp. 619–639, Oct. 2017.

[23] A. Samat, C. Persello, S. Liu, E. Li, Z. Miao, and J. Abuduwaili, "Classification of VHR multispectral images using extratrees and maximally stable extremal region-guided morphological profile," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 9, pp. 3179–3195, Sep. 2018.

[24] A. Samat, E. Li, P. Du, S. Liu, Z. Miao, and W. Zhang, "CatBoost for RS image classification with pseudo label support from neighbor patches-based clustering," IEEE Geosci. Remote Sens. Lett., to be published, doi: 10.1109/LGRS.2020.3038771.

[25] A. Samat, E. Li, W. Wang, S. Liu, C. Lin, and J. Abuduwaili, "Meta-XGBoost for hyperspectral image classification using extended MSER-Guided morphological profiles," Remote Sens., vol. 12, no. 12, Jun. 2020, Art. no. 1973.

[26] R. Liu, W. Li, X. Liu, X. Lu, T. Li, and Q. Guo, "An ensemble of classifiers based on positive and unlabeled data in one-class remote sensing classification," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 2, pp. 572–584, Feb. 2018.

[27] X. Li, X. Liu, and L. Yu, "Aggregative model-based classifier ensemble for improving land-use/cover classification of landsat TM images," Int. J. Remote Sens., vol. 35, no. 4, pp. 1481–1495, Feb. 2014.

[28] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[29] H. Shen, Y. Lin, Q. Tian, K. Xu, and J. Jiao, "A comparison of multiple classifier combinations using different voting-weights for remote sensing image classification," Int. J. Remote Sens., vol. 39, no. 11, pp. 3705–3722, Mar. 2018.

[30] J. Chen, J. Xia, P. Du, and J. Chanussot, "Combining rotation forest and multiscale segmentation for the classification of hyperspectral data," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 9, no. 9, pp. 4060–4072, Sep. 2016.

[31] T. Berhane et al., "Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory," Remote Sens., vol. 10, no. 4, Apr. 2018, Art. no. 580.

[32] J. Xiao, H. Wu, C. Wang, and H. Xia, "Land cover classification using features generated from annual time-series landsat data," IEEE Geosci. Remote Sens. Lett., vol. 15, no. 5, pp. 739–743, May 2018.

[33] P. Dou and Y. Chen, "Remote sensing imagery classification using adaboost with a weight vector (WV adaboost)," Remote Sens. Lett., vol. 8, no. 8, pp. 733–742, 2017.

[34] S. Liu and Q. Shi, "Local climate zone mapping as remote sensing scene classification using deep learning: A case study of metropolitan China," *ISPRS J. Photogramm. Remote Sens.*, vol. 164, pp. 229–242, Jun. 2020.

[35] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[36] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3014312.

[37] M. Wang, X. Zhang, X. Niu, F. Wang, and X. Zhang, "Scene classification of high-resolution remotely sensed image based on Resnet," *J Geovis. Spatial Anal.*, vol. 3, no. 2, Oct. 2019, Art. no. 16.

[38] P. Dou and C. Zeng, "Hyperspectral image classification using feature relations map learning," *Remote Sens.*, vol. 12, no. 18, Sep. 2020, Art. no. 2956.

[39] Y. Su, X. Xu, J. Li, H. Qi, P. Gamba, and A. Plaza, "Deep autoencoders with multitask learning for bilinear hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3041157.

[40] Y. Zhong, F. Fei, Y. Liu, B. Zhao, H. Jiao, and L. Zhang, "SatCNN: Satellite image dataset classification using agile convolutional neural networks," *Remote Sens. Lett.*, vol. 8, no. 2, pp. 136–145, Oct. 2016.

[41] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, Sep. 2018.

[42] H. Lyu *et al.*, "Long-term annual mapping of four cities on different continents by applying a deep information learning method to landsat data," *Remote Sens.*, vol. 10, no. 3, Mar. 2018, Art. no. 471.

[43] M. A. Dede, E. Aptoula, and Y. Genc, "Deep network ensembles for aerial scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 732–735, May 2019.

[44] Y. Chen, Y. Wang, Y. Gu, X. He, P. Ghamisi, and X. Jia, "Deep learning ensemble for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1882–1897, Jun. 2019.

[45] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[46] P. Zhong, Z. Gong, S. Li, and C.-B. Schonlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.

[47] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[48] Q. Weng, Z. Mao, J. Lin, and X. Liao, "Land-use scene classification based on a CNN using a constrained extreme learning machine," *Int. J. Remote Sens.*, vol. 39, no. 19, pp. 6281–6299, Apr. 2018.

[49] C. Zhang *et al.*, "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, Feb. 2019.

[50] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1685–1689, Oct. 2017.

[51] S. Xie, L. Liu, X. Zhang, and X. Chen, "Annual land-cover mapping based on multi-temporal cloud-contaminated landsat images," *Int. J. Remote Sens.*, vol. 40, no. 10, pp. 3855–3877, Dec. 2018.

[52] X.-C. Yin, C. Yang, W.-Y. Pei, and H.-W. Hao, "Shallow classification or deep learning: An experimental study," in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 1904–1909.

[53] G. Ososkov and P. Goncharov, "Shallow and deep learning for image classification," *Opt. Memory Neural Netw.*, vol. 26, no. 4, pp. 221–248, Oct. 2017.

Huanfeng Shen (Senior Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

In 2007, he joined the School of Resource and Environmental Sciences (SRES), Wuhan University, where he is currently a Luojia Distinguished Professor and an Associate Dean with SRES. He was or is the PI of two projects supported by National Key Research and Development Program of China, and six projects supported by National Natural Science Foundation of China. He has authored more than 100 research papers in peer-reviewed international journals. His research interests include remote sensing image processing, multisource data fusion, and intelligent environmental sensing.

Dr. Shen is a Council Member of China Association of Remote Sensing Application, Education Committee Member of Chinese Society for Geodesy Photogrammetry and Cartography, and Theory Committee Member of Chinese Society for Geospatial Information Society. He is currently a member of the editorial board of *Journal of Applied Remote Sensing* and *Geography and Geo-Information Science*.

Zhiwei Li received the B.S. degree in geoinformation science and technology from China University of Geosciences, Wuhan, China, in 2015, and the Ph.D. degree in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2020.
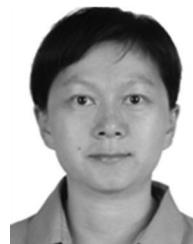
He is currently a Postdoctoral Researcher with the School of Resource and Environmental Sciences, Wuhan University. His research interests include spatial–temporal seamless reconstruction of remote sensing data, land cover/use classification, and urban remote sensing mapping.
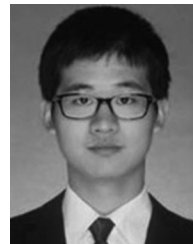
Xiaobin Guan received the B.S. and Ph.D. degrees in geographical information system from the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, in 2013 and 2018, respectively.

He is currently a Postdoctoral Research Assistant with the School of Resource and Environmental Sciences, Wuhan University. His research interests include the processing of multisource remote-sensing images and its application in the terrestrial ecosystem and global change.

Peng Dou received the M.S. degree in cartography and geography information system from Lanzhoujiaotong University, Lanzhou, China, in 2014, the Ph.D. degree in cartography and geography information system from Sun Yat-Sen University, Guangzhou, China, in 2018.

He is currently a Postdoctoral Researcher with the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China. His research interests include remote sensing image classification, urbanization dynamic detection, and image processing.

Wenli Huang (Member, IEEE) received the B.S. degree in geographic information system (computer cartography) from Wuhan University, Wuhan, China, in 2006, the M.S. degree in cartography and geographic information system from Beijing Normal University, Beijing, China, in 2009, and the Ph.D. degree in geographical sciences from the University of Maryland, College Park, MD, USA, in 2015.

From 2015 to 2018, she was a Postdoctoral Research Associate of the geographical sciences with the University of Maryland. Since 2019, she has been an Associate Research Professor with the School of Resource and Environmental Sciences, Wuhan University. Her research interests include mapping of land cover and land use, forest canopy structures, and aboveground biomass through remote sensing.