IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 4, 2020, accepted March 27, 2020, date of publication April 7, 2020, date of current version April 29, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986267

Small Sample Classification of Hyperspectral Remote Sensing Images Based on Sequential Joint Deeping Learning Model

ZESONG WANG[1], CUI ZOU[1], AND WEIWEI CAI[2], (Graduate Student Member, IEEE)

[1]The Big Data Institute, Qingdao Huanghai University, Qingdao 266427, China
[2]School of Logistics and Transportation, Central South University of Forestry and Technology, Changsha 410004, China

Corresponding author: Zesong Wang (qdhhxyw@163.com)

This work was supported in part by the Key Research and Development plan of Shandong Province under Grant 2019GGX105001, and in part by 2019 Shandong province colleges and universities young talents introduction plan construction team project: big data and business intelligence social service innovation team.

ABSTRACT Although hyperspectral remote sensing images have rich spectral features, for small samples of remote sensing images, feature selection, feature mining, and feature integration are very important. A single model is difficult to apply to multiple tasks such as feature selection, feature mining, and feature integration during training, resulting in poor classification results for small sample classification of hyperspectral images. To improve the classification of small samples, a sequential joint deep learning algorithm is proposed in this paper. (In this algorithm, the deep features of multiscale convolution under an attention mechanism are integrated by using Bidirectional Long Short-Term Memory(Bi-LSTM) and AML.) First, we used principal component analysis (PCA) to reduce the dimensionality of the hyperspectral data and retain their key features. Second, the model uses an integrated attention mechanism to distribute the probability weight of the key input feature. Third, the model uses multiscale convolution to mine features after the distribution weight to obtain deep features. Fourth, the model uses bidirectional long short-term memory (Bi-LSTM) to integrate the convolution results at different scales. Finally, the softmax classifier is used to complete the classification of multiclass hyperspectral remote sensing images. Experiments were carried out on three public hyperspectral data sets, and the results proved that our proposed AML algorithm is effective, thus demonstrating powerful performance in the prediction of hyperspectral images (HSIs) of small samples.

INDEX TERMS Integrated attention mechanism, multi-scale convolution operation, features fusion, HSIs.

I. INTRODUCTION

Hyperspectral images (HSIs) contain hundreds of bands in each pixel. Due to the richness of the hyperspectral image spectrum, HSIs are widely used in agriculture [1], forestry [2], and urban topography [3]. However, the rich bands bring sufficient features to HSIs while also producing many redundant features. Recently, researchers have proposed new methods for solving redundant features and choosing useful features. Nie *et al.* [9] proposed a model for the automatic weighting of features to minimize redundant features. By establishing relative cosine distances for different redundant features, Ayinde *et al.* [10] ultimately eliminated many redundant features and reduced model computational costs. We chose principal components analysis (PCA) because it is important in eliminating redundant features [11]–[14] and is not subject to parameter settings. Small samples of HSIs are also a difficult problem in feature mining and classification. To improve the classification of hyperspectral images of small samples, the researchers have investigated several methods, such as support vector machines (SVMs) [4], conditional random fields (CRF) [5], k-nearest neighbors (KNN) algorithms [15], and clustering-based classification [16]. However, the above methods are insufficient for mining hyperspectral features or eliminating more redundant features. Recently, increasingly greater achievements have been made in image classification by deep learning due to its powerful feature learning and classification ability [17]. To improve classification, Mesut *et al.* [6] used AlexNet to extract morphological features of hyperspectral images and enhance the mining of spatial features through morphological features.

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

To eliminate the redundancy and noise features of hyperspectral images, Mercedes E. *et al.* [7] proposed a residual neural network (ResNet) to gradually increase the dimension of convolution and, finally, eliminate redundant features. Yang *et al.* [8] proposed a dual-channel dense convolutional network (DenseNet) to extract the spectral and spatial features of hyperspectral images and further improved the accuracy of the classification of hyperspectral images by fully using useful features. These methods have performed excellently in feature mining. By using a long short-term memory (LSTM) algorithm, Zhoucheng *et al.* [19] proved that the long-term sequence properties of features in hyperspectral images play a positive role in classification.

However, due to the nature of the convolution kernel size, the final learning features are determined. Hyperspectral images have many features and are difficult to adapt to a single convolution kernel size. As the number of model layers is increased, increasingly more useful features are lost. For small samples and high-dimensional hyperspectral images, it becomes more difficult to learn complete image features; consequently, it becomes difficult to accurately identify complex hyperspectral images. H Lee and H Kwon [18] proposed a contextual CNN that achieved good results in excavating hyperspectral image features of different scales. However, feature mining has only been partly improved, and methods of feature selection and feature integration are still insufficiently comprehensive. To solve the above problems, we propose a sequential joint deep learning algorithm [27](The Deep features of Multi-scale convolution under Attention Mechanism are integrated by Bi-LSTM, AML). This algorithm plays an active role in feature selection, feature mining and feature integration. The experimental results show that the proposed algorithm can better select, mine and integrate the features of hyperspectral images.

The main contributions of this paper are as follows.

- We propose an integrated attention mechanism to distribute the probability weight of the key feature, thereby enhancing the selection of features in the model.
- We propose a multiscale convolution algorithm to retain more deep features, thereby enhancing the feature-mining ability of the model.
- To better integrate deep features, we use bidirectional long short-term memory (Bi-LSTM) [21] to integrate the deep features of convolution kernel mining at each scale.
- To better understand the training results of the AML algorithm, convergence visualization analysis was performed on three public datasets.

The rest of this paper is summarized as follows: Section II introduces the AML algorithm. Section III analyzes the experimental results. Finally, the conclusion is drawn in Section IV.

II. RELATED RESEARCH

An excellent classification model performs three tasks: feature selection, feature mining and feature integration; additionally, an excellent classification model is important in the final hyperspectral remote sensing image classification. First, the selection of good features provides a guarantee for later feature mining and feature integration. Jie [22] and others set 11 filters in the first layer of depth convolution to select bands. Although many redundant bands are eliminated, the useful features in bands are simultaneously lost, and the final classification effect is reduced. Ying [23] and others trained the hyperspectral data through a self-defined one-dimensional CNN and predicted the data in other bands. Finally, the authors considered the band with the highest accuracy as the selected band, tested all the bands repeatedly, and finally achieved satisfactory results on the Indian Pines dataset. Hongfeng [27] proposed a SAGP algorithm, which firstly uses the attention mechanism to extract features from remote sensing images, allocates weight coefficients for features, and preserves the integrity of features for later feature mining, although the attention mechanism assigns weights to the original features, it ignores the weight distribution of global and local features. The PRAN algorithm proposed by Gao H [29] combines the advantages of an attention mechanism model and a residual network, jointly constructs spectral and spatial features, and improves the robustness of learning, especially on small sample training sets. Then, a deep mining of features is needed after feature selection. Jie [24] uses a multilayer CNN that improves classification accuracy by extracting shallow, middle and deep features of hyperspectral remote sensing images and fusing a variety of features to make shallow features fully complementary to deep features. Chunju [25] uses multiscale convolution to mine multiscale features, integrates the characteristic information of the whole network mining, and achieves high-precision classification results and efficient operation speed on multiple data sets. The final task performed by the model is feature integration. Lefei *et al.* [30] map spatial and spectral features to a common low-dimensional subspace, and many kinds of feature information complement each other. Additionally, the authors use the model to mine the most important original image features, which are proved on three hyperspectral remote sensing data sets. Ji C *et al.* [31] proposed an NMF depth feature extraction algorithm, which transmits features layer by layer by iterating several nonnegative matrices and reconstructing the residual network. Additionally, the activation function is used to enhance the ability of nonlinear feature extraction. The results show that the algorithm is efficient and suitable for hyperspectral remote sensing image classification. After fully mining deep features, feature integration can better retain the relationship between sequence structure information and constructed features. Lichao [26] proved that hyperspectral images essentially have a sequence-based data structure and proposed a new RNN model algorithm to analyze the sequence data in hyperspectral images more effectively; finally, the authors proved that their depth recursive network has great potential in hyperspectral data. Chen J [28] and others proposed an FSSFNet algorithm that achieved good experimental results

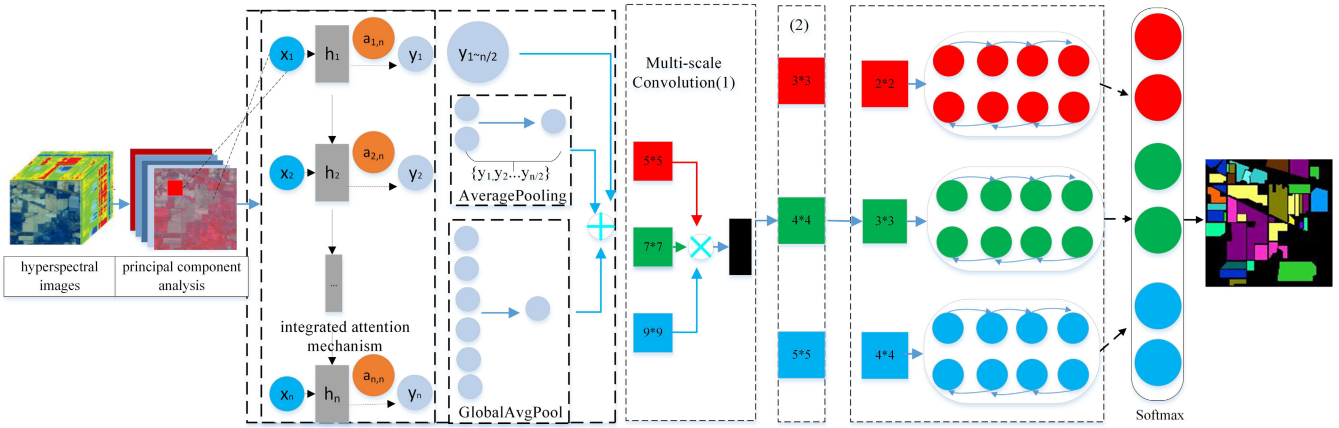

FIGURE 1. AML algorithm model. The numbers in the square represent the size of the convolution kernel. Circles represent the eigenvalues of each part of the model. ⊗ represents the feature fusion operation. The black rectangle represents the max-pooling layer. (1) and (2) represent two multiscale convolutions. ⊕ represents the feature sum operation.

by, first, eliminating the redundant features in multispectral features; then, constructing the spatial relationship features between pixels; and, finally, combining these features.

The abovementioned research clearly shows that feature selection, feature mining and feature integration are indispensable. In this paper, in the feature selection stage, we propose an integrated attention mechanism model to select the original features of hyperspectral remote sensing images. The advantage of this model is that it assigns a higher weight coefficient to the recognition of target features so that the target features can be better mined, thereby avoiding the elimination of the whole band with less information. The model simultaneously combines local and global feature weight information. In the feature mining stage, we select the multiscale convolution model to fully mine the features of different scales and retain more useful features from multiple scales. In the feature integration stage, we use Bi-LSTM to integrate the deep features of each kind of convolution mining and retain the sequence relationship between the deep features of each kind of convolution mining. Experimental results show that our proposed algorithm plays a positive role in feature selection, feature mining and feature integration.

III. PROPOSED FRAMEWORK

Figure 1 shows the AML algorithm model. First, the original hyperspectral image is reduced by the PCA algorithm, which retains most of the information dimensions and reduces the computational cost. Second, the dimension-reduced features are input into the AML algorithm. In the AML algorithm, the attention mechanism is used to distribute the weight of the input features, then the output features of attention are selected by three processing methods, and then the calculated features are input to multi-scale convolution for deep mining. Finally, the mined features are put into BILSTM for final feature integration. The final integrated features are predicted by softmax, and realize the recognition of hyperspectral images. The three attention mechanisms not only retain the original features of high-dimensional images, but also add local and global features. In the next few parts of this chapter, we introduce each part of the AML algorithm in detail.

A. INTEGRATED ATTENTION MECHANISM

An attention mechanism [20] is a type of probability weight distribution mechanism. By calculating the features of remote sensing images input at different times, this mechanism pays more attention to the features of target recognition. Thus, these features are assigned larger probability weights, which contribute to improving the quality of the hidden-layer features. The weight coefficient of the hidden layer is calculated as follows (Equation 1):

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{n} exp(e_{ik})} \tag{1}$$

where a_{ij} represents the attention allocation coefficient of the feature sequence of the pixel j in a source input, i represent a moment, e represents the energy value of the i-th moment.

$$c_i = \sum_{j=1}^{n} \alpha_{ij} h_j \tag{2}$$

where h_j represents the hidden state information of the pixel j by feature vectors. Finally, we used summation functions for a_{ij} and the hidden state h_j to generate the context vector c_i.

The integrated attention mechanism we propose consists of three parts. In the first part, we retain half of the hyperspectral image features with high weight; in the second part, we use the local average pooling strategy to pool the adjacent features and retain the number of new local average pooled hyperspectral image features to half of the original hyperspectral image features; in the third part, we use the global average pooling strategy to average the overall hyperspectral image features and generate a new global average pooling hyperspectral image feature. By using the above three attention

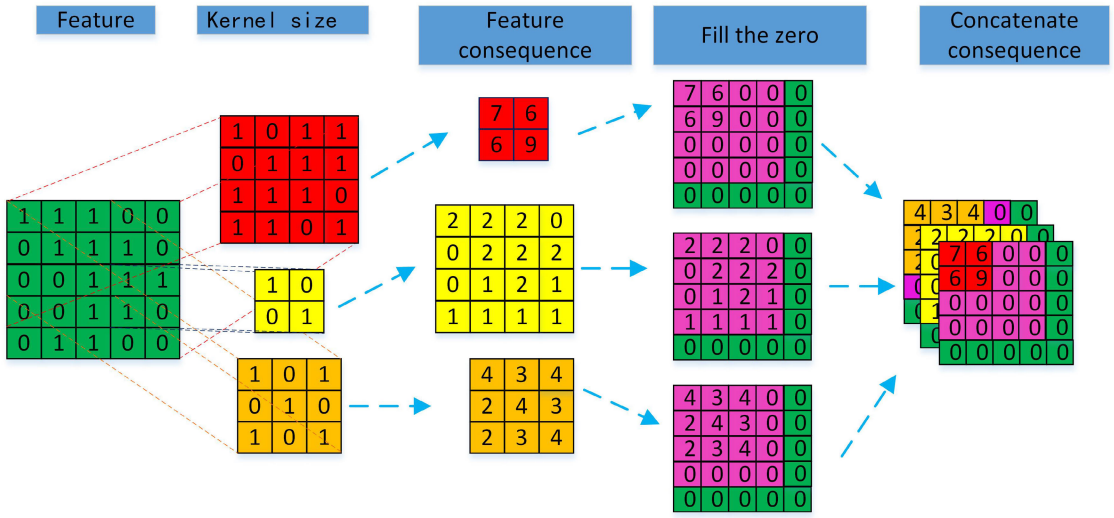

FIGURE 2. The multiscale convolution operation. Green indicates the input features; red, orange and yellow indicate the size of the convolution kernel at different scales, and the pink matrix pads the red, orange and yellow matrices. The multicolored matrix indicates the features result of the output. The multiscale feature is fused by 'concatenate' and is used as the input of the next module. Finally, the output feature map is kept the same size as the input image.

processing methods, we finally extract the original, neighborhood, and global features of the hyperspectral image. Using these methods provides a guarantee for feature mining and feature integration. The specific operation is shown in the integrated attention mechanism module in Figure 1.

B. MULTISCALE CONVOLUTION

The features are directly affected by the size of the convolution kernel. Traditional convolution works only through a convolution kernel, and a corresponding basic feature is obtained. Sometimes, the single convolution kernel we choose does not fully exploit the useful features in the image, thus resulting in the loss of some key features.

To solve the above problems, we propose a novel multiscale convolution method, which simultaneously mines features from multiple scales through multiple convolution kernels. The experimental results show that multiscale convolution can better preserve key features. The multiscale convolution operation is illustrated in Figure 2.

The multiscale convolution formula is as follows:

$$S = Conv \sum_{k=1}^{n} (X_{(i,j)} * W_{(k-1,k,k+2)}) + b_{(k-1,k,k+1)} \quad (3)$$

where $W_{(k-1,k,k+2)}$ represents the weight calculation with the k−1, k, k+1 convolution kernel. $b_{(k-1,k,k+1)}$ represents a bias function that simultaneously adds the k−1, k, k+1 convolution kernel.

C. BIDIRECTIONAL LONG SHORT-TERM MEMORY

We integrated the deep features generated by the last layer of multiscale convolution mining to integrate their own information. We improve the correlation between deep features through Bi-LSTM. In this paper, three Bi-LSTM layers are used to integrate the upper and lower semantic information of the deep features of three different sizes of convolution to provide more contextual semantic information features for the final classification of small samples. Additionally, Bi-LSTM effectively avoids the problem of gradient explosion and gradient disappearance. Bi-LSTM consists of a forward LSTM called $LSTM_{forward}$ and a backward LSTM called $LSTM_{backward}$. Bi-LSTM uses $LSTM_{forward}$ and $LSTM_{backward}$ to strengthen the context semantic information between deep features to complete deep feature integration. The visualization of the integration is shown in Figure 3.

The forward and backward propagation fusion formula of Bi-LSTM is as follows:

$$Integ = LSTM_{forward} \oplus LSTM_{backward} \quad (4)$$

where *Integ* represents the forward and backward propagation integrated output of Bi-LSTM. $\oplus$ represents an integrated operational symbol.

D. FEATURE FUSION

We fused three Bi-LSTM integrated deep feature sequences. The fusion algorithm formula is as follows:

$$Output = Concatenate(Integ_1 \oplus Integ_2 \oplus Integ_3) \quad (5)$$

where *Output* represents the output of the final merged feature fusion; *Concatenate()* represents a fusion feature function; $Integ_1$ represents a deep feature sequence generated by Bi-LSTM with a convolution kernel scale of k-1; $Integ_2$ represents a deep feature sequence generated by Bi-LSTM with a convolution kernel scale of k; $Integ_3$ represents a deep feature sequence generated by Bi-LSTM with a convolution kernel scale of k+1. The fusion features are finally classified by softmax function.

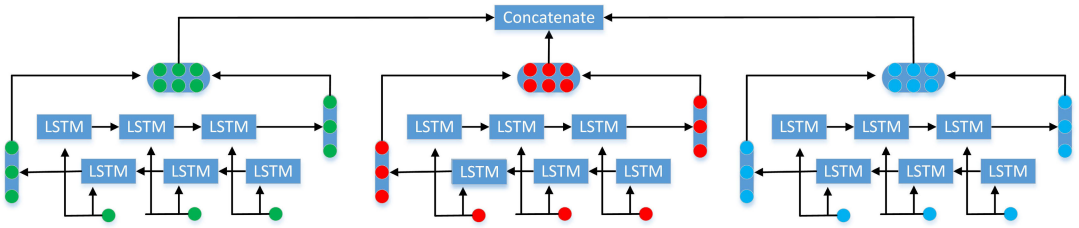

FIGURE 3. The feature integration model picture representing the generation of multiscale Bi-LSTM. Each red circle represents a feature with a mining scale equal to 2. Each green circle represents a feature with a mining scale equal to 3. Each blue circle represents a feature with a mining scale equal to 4. Concatenate represents the integrated operation of features.

TABLE 1. The number of samples of each class in the Indina Pines dataset.

Number	Class	Samples
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

TABLE 2. The number of samples of each class in the Pavia University scene dataset.

Number	Class	Samples
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947

TABLE 3. The number of samples of each class in the Salinas dataset.

Number	Class	Samples
1	Brocoli-green-weeds-1	2009
2	Brocoli-green-weeds-2	3726
3	Fallow	1976
4	Fallow-rough-plow	1394
5	Fallow-smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes-untrained	11271
9	Soil-vinyard-develop	6203
10	Corn-senesced-green-weeds	3278
11	Lettuce-romaine-4wk	1068
12	Lettuce-romaine-5wk	1927
13	Lettuce-romaine-6wk	916
14	Lettuce-romaine-7wk	1070
15	Vinyard-untrained	7268
16	Vinyard-vertical-trellis	1807

IV. EXPERIMENTS

A. DATA SETS AND EVALUATION METHODS

In this paper, our experiments were conducted based on three hyperspectral public data sets. The first data set is Indina Pines, which is 145*145 in size and reduced to 100-D via PCA. The entire image has 16 different classes. The second data set is the Pavia University scene, which is 610*340 in size and reduced to 50-D via PCA. The entire image has 9 different classes. The third data set is Salinas, which is 512*217 in size and reduced to 29-D via PCA. The entire image has 16 different classes. Overall accuracy (OA), average accuracy (AA), and the kappa coefficient (kappa) are used as the performance metrics to evaluate the classification accuracy of the letter. To ensure the validity of the experimental data, the experiment is repeated 10 times, the highest value is taken, and each test randomly selected the samples. We describe the number of samples of the three data sets in detail in Table 1, Table 2, Table 3.

B. EXPERIMENTAL RESULTS OF DIFFERENT METHODS

We compare the proposed AML algorithm with the classical AlexNet, ResNet, and DenseNet algorithms and the latest FSSFNet, SAGP, and PRAN algorithms. To ensure fairness, all models use the same features after PCA dimensionality reduction and the same number of samples in the training and verification set. The results are shown in the following tables for the different data sets (Table 4 for the Indian Pine dataset; Table 5 for the Pavia University dataset and Table 6 for the Salinas dataset).

1) RESULTS ON THE INDIAN PINE DataSet

The Indian Pines data set belongs to a small sample data set, of which 9 are less than 500 samples and the smallest class has only 20 samples. Table 4 clearly shows that the effect of our proposed AML operation is higher than that of other contrast algorithms by 1% to 8% in terms of OA, AA and Kappa. Our proposed model shows the best results on three different

TABLE 4. Classification performance of different approaches for the Indian Pine image with 5%, 10%, AND 15% training samples. Bold indicates the result of the improved algorithm.

Methods	OA(5%)	AA(5%)	Kappa(5%)	OA(10%)	AA(10%)	Kappa(10%)	OA(15%)	AA(15%)	Kappa(15%)
AlexNet [6]	0.6896	0.5692	0.6450	0.7423	0.6571	0.7049	0.8167	0.7955	0.7907
ResNet [7]	0.7076	0.6996	0.6640	0.7804	0.7911	0.7501	0.8300	0.8067	0.8059
DenseNet [8]	0.7114	0.6732	0.6683	0.7892	0.7551	0.7579	0.8428	0.8149	0.8203
PRAN [29]	0.7276	0.7359	0.6975	0.7782	0.7343	0.7459	0.8286	0.7663	0.8039
FSSFNet [28]	0.7375	0.6795	0.6998	0.7866	0.7126	0.7547	0.8261	0.7448	0.8005
SAGP [27]	0.7349	0.7658	0.7361	0.7836	0.8089	0.7672	0.8159	0.8650	0.8289
AML	**0.7704**	**0.7772**	**0.7446**	**0.8295**	**0.8329**	**0.8189**	**0.8825**	**0.8948**	**0.8650**

TABLE 5. Classification performance of different approaches for the Pavia university image with 1%, 5%, AND 10% training samples. Bold indicates the result of the improved algorithm.

Methods	OA(1%)	AA(1%)	Kappa(1%)	OA(5%)	AA(5%)	Kappa(5%)	OA(10%)	AA(10%)	Kappa(10%)
AlexNet [6]	0.8706	0.8439	0.8271	0.9290	0.9182	0.9056	0.9298	0.9359	0.9149
ResNet [7]	0.8418	0.8245	0.7761	0.9090	0.9142	0.8785	0.9424	0.9489	0.9322
DenseNet [8]	0.8248	0.8075	0.7572	0.9036	0.8951	0.8719	0.9239	0.9355	0.9145
PRAN [29]	0.8938	0.8872	0.8574	0.9359	0.9247	0.9151	0.9482	0.9362	0.9314
FSSFNet [28]	0.8663	0.8423	0.8211	0.9372	0.9235	0.9165	0.9437	0.9265	0.9252
SAGP [27]	0.8432	0.8384	0.7883	0.9109	0.8975	0.8814	0.9373	0.9297	0.9165
AML	**0.8981**	**0.8951**	**0.8632**	**0.9425**	**0.9328**	**0.9237**	**0.9543**	**0.9483**	**0.9392**

TABLE 6. Classification performance of different approaches for the Salinas image with 1%, 5%, and 10% training samples. Bold indicates the result of the improved algorithm.

Methods	OA(1%)	AA(1%)	Kappa(1%)	OA(5%)	AA(5%)	Kappa(5%)	OA(10%)	AA(10%)	Kappa(10%)
AlexNet [6]	0.9091	0.9384	0.8987	0.9419	0.9676	0.9353	0.9437	0.9579	0.9531
ResNet [7]	0.8739	0.9212	0.8596	0.9180	0.9529	0.9087	0.9379	0.9685	0.9309
DenseNet [8]	0.8565	0.9097	0.8404	0.9114	0.9447	0.9012	0.9309	0.9605	0.9229
PRAN [29]	0.7898	0.7622	0.7655	0.9033	0.8823	0.8922	0.9195	0.8944	0.9102
FSSFNet [28]	0.9104	0.9474	0.9002	0.9356	0.9669	0.9282	0.9585	**0.9803**	0.9537
SAGP [27]	0.9087	0.9468	0.8982	0.9264	0.9569	0.9181	0.9476	0.9745	0.9416
AML	**0.9163**	**0.9458**	**0.9315**	**0.9454**	**0.9715**	**0.9552**	**0.9593**	0.9780	**0.9666**

training sets, thus fully proving the feasibility of the AML algorithm for small sample detection. In SAGP, the sequence model is used to construct the sequence relationship between features so that a few samples can obtain better results. The AML algorithm plays an active role in feature selection, feature mining and feature fusion. Figure 4 shows a graph of the experimental results (generated using by 5% of the samples for training). The graph clearly shows that the AML model significantly reduces the noise generated by the classification effect.

2) RESULTS ON THE PAVIA UNIVERSITY DataSet

These data is divided into 9 categories; each class has thousands of samples, but the shape of the samples is irregular. Table 5 clearly shows that our AML algorithm improves OA, AA, and Kappa, and the evaluation criteria increase significantly as the number of training samples increases. In the PRAN model, good results are obtained by using two attention mechanism modules. Figure 5 shows the experimental results (generated using by 1% of the samples for training). This figure clearly shows that as the proportion of training samples increases, the noise significantly decreases. The classification effect of the AML algorithm is obviously stronger due to the traditional algorithm. (The red circle shows the obvious difference in some places between the original algorithm and our improved algorithm.)

3) RESULTS ON THE SALINAS DataSet

Salinas is a morphological relative rule data set, and the number of samples per class is large. Therefore, Table 6 shows that even if the training set accounts for only 1%, a better effect can be trained. Our improved ALM algorithm has obtained the best classification results. When the training set sample is 10%, the AA prediction result of FSSFNet is higher than that of our proposed model, thus showing that the global information in small samples is very important. Figure 6 shows the experimental results (generated using by 1% of the samples for training). Figure 6 shows very clearly that our proposed AML algorithm has stronger feature learning ability, thus improving the classification accuracy of small samples.

C. EXPERIMENTAL RESULTS OF DIFFERENT ATTENTIONAL MECHANISMS

In this section, we mainly compare the effects of various attention mechanisms on the experimental results. We used

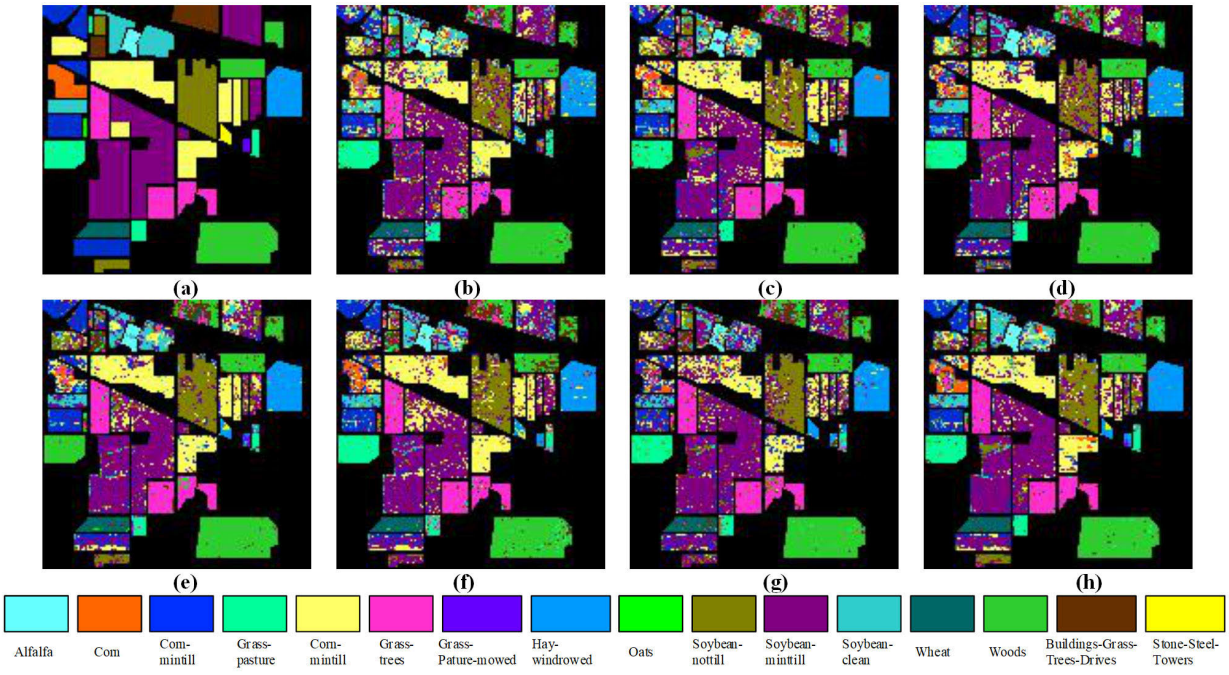

FIGURE 4. Classification maps of the Indian pine data set. (a) Ground truth. (b) AlexNet. (c) ResNet. (d) DenseNet. (e) PRAN. (f) FSSFNet. (g) SAGP. (h) AML.

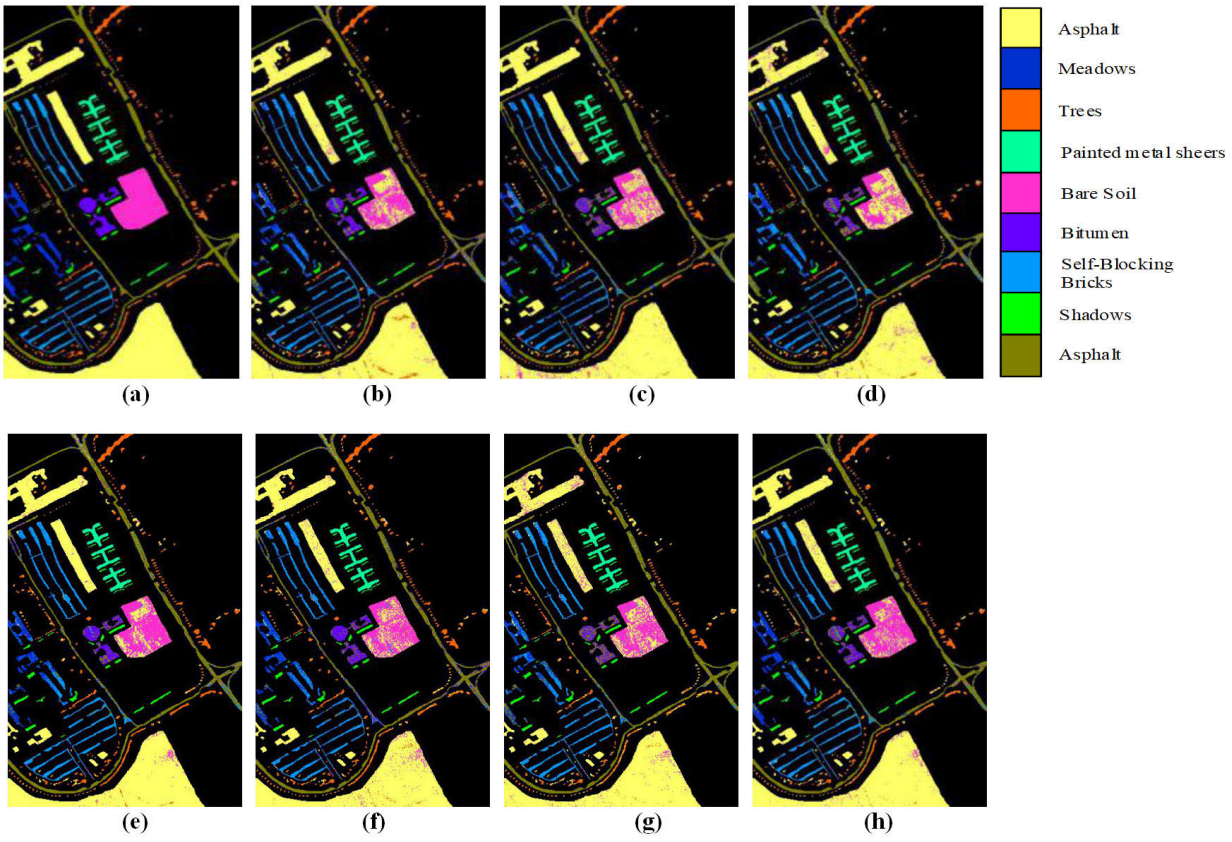

FIGURE 5. Classification maps of the University of Pavia data set. (a) Ground truth. (b) AlexNet. (c) ResNet. (d) DenseNet. (e) PRAN. (f) FSSFNet. (g) SAGP. (h) AML.

the least divided training samples for each dataset for testing (the Indian pine data set used 5%, the University of Pavia data set used 1%, and the Salinas data set used 1%). The number "1" represents the traditional attention mechanism [20]. "2" represents the attention mechanism of average pooling operation. and "3" represents the attention

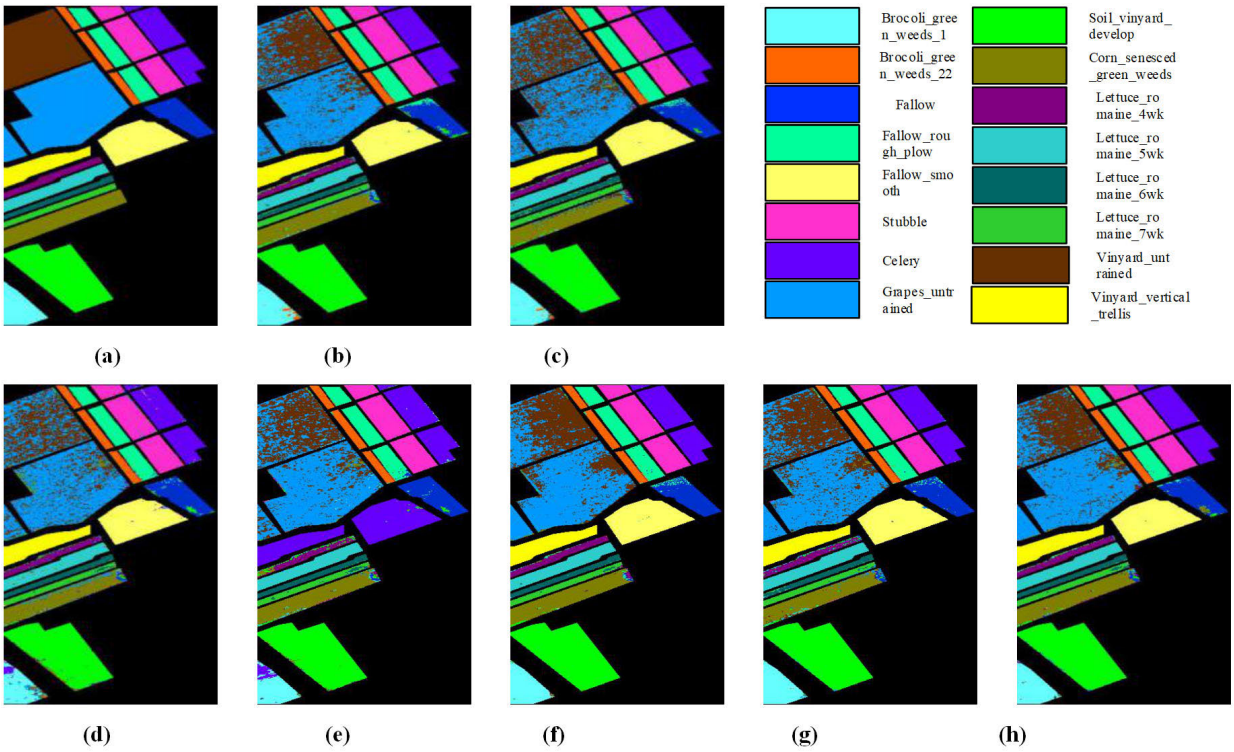

FIGURE 6. Classification maps of the Salinas data set. (a) Ground truth. (b) AlexNet. (c) ResNet. (d) DenseNet. (e) PRAN. (f) FSSFNet. (g) SAGP. (h) AML.

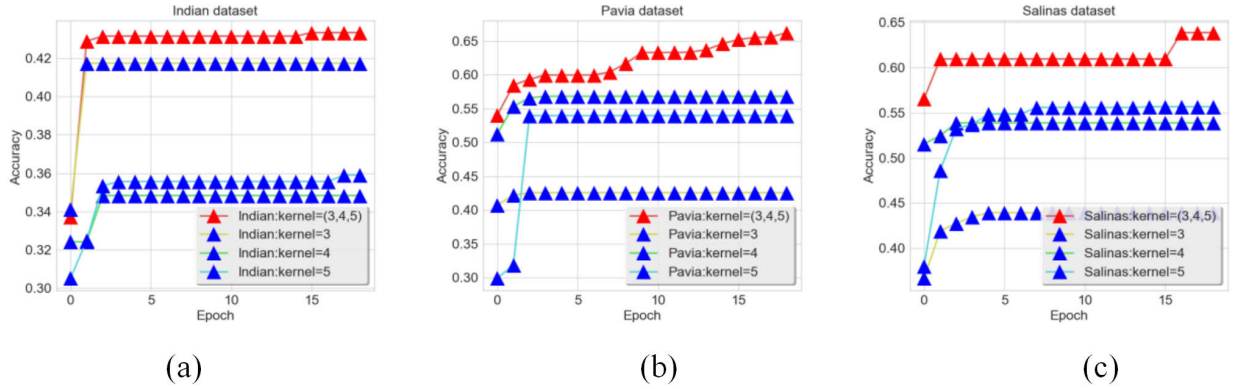

FIGURE 7. Convergence curves of each dataset. (a) Indian data. (b) Pavia data. (c) Salinas data. Red is multi-scale convolution, blue is traditional convolution.

TABLE 7. Experimental results of different attention mechanisms.

Methods	OA(IP)	AA(IP)	Kappa(IP)	OA(UA)	AA(UA)	Kappa(UA)	OA(Salinas)	AA(Salinas)	Kappa(Salinas)
AML-no-Attention	0.7451	0.7449	0.7198	0.8795	0.8673	0.8380	0.8930	0.9269	0.8809
AML-Attention-1	0.7556	0.7618	0.7231	0.8882	0.8692	0.8501	0.9033	0.9272	0.8923
AML-Attention-2	0.7502	0.7627	0.7272	0.8870	0.8710	0.8485	0.9032	0.9248	0.8921
AML-Attention-1-2	0.7581	0.7702	0.7362	0.8942	0.8937	0.8605	0.9068	0.9331	0.8963
AML-Attention-1-3	0.7582	0.7725	0.7395	0.8922	0.8944	0.8551	0.9047	0.9286	0.8938
AML-Attention-2-3	0.7559	0.7659	0.7310	0.8895	0.8872	0.852	0.9073	0.9276	0.8974
AML	**0.7704**	**0.7772**	**0.7446**	**0.8997**	**0.8983**	**0.8652**	**0.9163**	**0.9458**	**0.9315**

mechanism of the global average pooling operation. Multiple numbers represent the simultaneous use of many different types of attention mechanisms, and the AML algorithm uses all the attention mechanisms. Global pooling attention generates only one feature, so the experiment is not compared separately. Table 7 clearly shows that the effect of any of the attention mechanism models is better than the effect of any of the models without an attention mechanism. The joint

TABLE 8. Experimental results of feature integration of different modules.

Methods	OA(IP)	AA(IP)	Kappa(IP)	OA(UA)	AA(UA)	Kappa(UA)	OA(Salinas)	AA(Salinas)	Kappa(Salinas)
AML-no-Bi-LSTM	0.7494	0.7518	0.7118	0.8673	0.8795	0.8380	0.8224	0.8161	0.8026
AML-GRU	0.7675	0.7728	0.7333	0.8963	0.8962	0.8611	0.8998	0.9363	0.8886
AML-Bi-GRU	0.7598	0.7619	0.7235	0.8729	0.8880	0.8495	0.8911	0.9281	0.8785
AML-LSTM	0.7686	0.7673	0.7252	0.8627	0.8778	0.8365	0.9009	0.9308	0.8895
AML	**0.7704**	**0.7772**	**0.7446**	**0.8997**	**0.8983**	**0.8652**	**0.9163**	**0.9458**	**0.9315**

TABLE 9. Experimental results of different submodules.

Methods	OA(IP)	AA(IP)	Kappa(IP)	OA(UA)	AA(UA)	Kappa(UA)	OA(Salinas)	AA(Salinas)	Kappa(Salinas)
AML-no-Attention	0.8663	0.8418	0.8476	0.9528	0.9420	0.9373	0.9206	0.8946	0.9115
AML-no-Multscale-Con	0.8434	0.8395	0.8214	0.9424	0.9369	0.9234	0.9581	0.9773	0.9534
AML-no-Bi-LSTM	0.8171	0.8215	0.7911	0.9536	0.9464	0.9384	0.9524	0.9746	0.9469
AML	**0.8825**	**0.8948**	**0.8650**	**0.9543**	**0.9483**	**0.9392**	**0.9593**	**0.9780**	**0.9666**

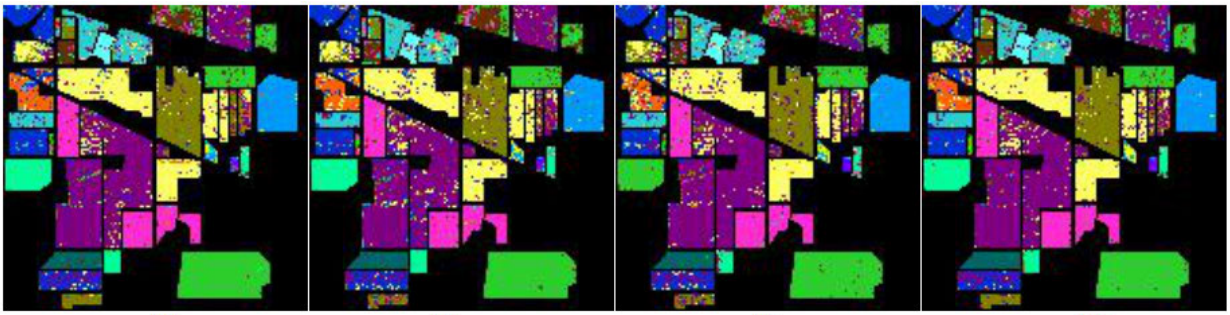

FIGURE 8. Classification maps of the Indian pine data set. (a) AML-no-Attention. (b) AML-no-Multscale-Con. (c) AML-no-Bi-LSTM. (d) AML.

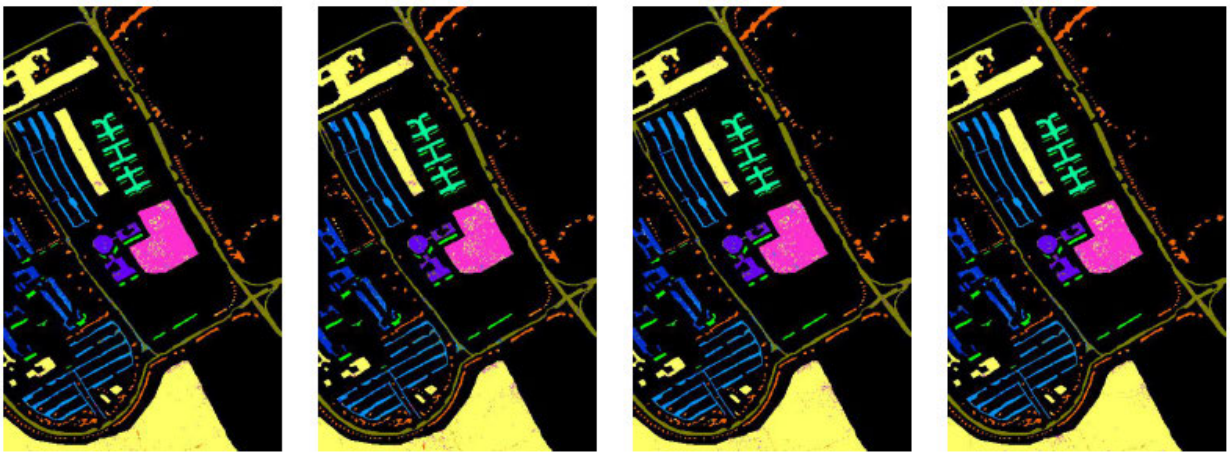

FIGURE 9. Classification maps of the University of Pavia data set. (a) AML-no-Attention. (b) AML-no-Multscale-Con. (c) AML-no-Bi-LSTM. (d) AML.

experimental results of the two attention mechanism models are higher than the experimental results of the single attention mechanism model. The traditional attention mechanism and the average pooling attention mechanism model are better than the other two combinations because there is only one feature of the global average pooling, so the generalization ability is limited. Our proposed AML algorithm optimizes the experimental results, so the three attention mechanism models are feasible.

D. COMPARISON OF TRADITIONAL CONVOLUTION AND MultiScale CONVOLUTION

To determine whether the traditional or multiscale convolution works better, we validated the convolutions on three datasets. We constructed a multiscale convolution set with convolution kernels being 3, 4, and 5. The experiment ran 20 rounds of multiscale and traditional convolution, 10 iterations per round. The experimental results are shown in Figure 7. The product verification of the multiscale

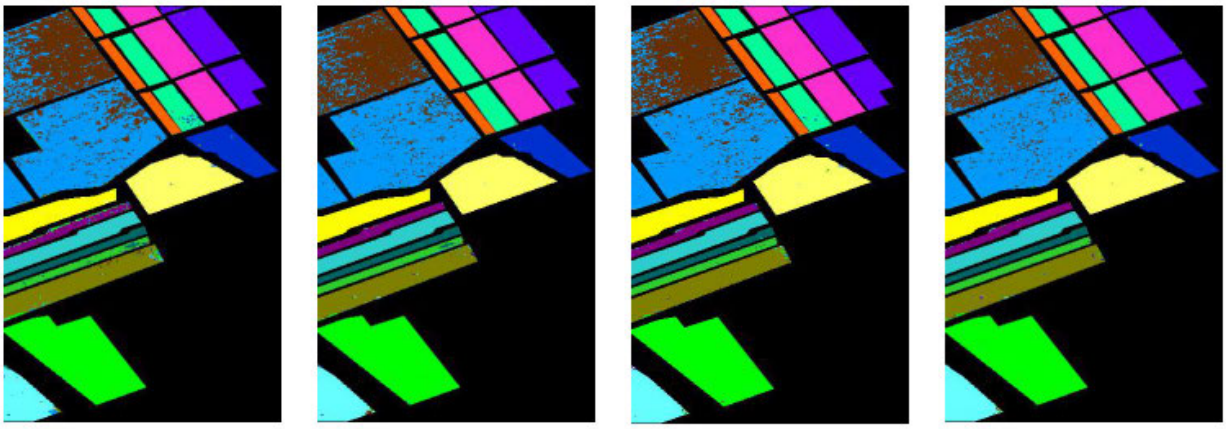

FIGURE 10. Classification maps of the Salinas data set. (a) AML-no-Attention. (b) AML-no-Multscale-Con. (c) AML-no-Bi-LSTM. (d) AML.

convolution module in Figure 7 comes from the second multiscale convolution module of our proposed model. The training set used by the Indian Pines dataset is 5% of its total data; the training set used by the Pavia dataset is 1% of its total data; and the training set used by the Salinas dataset is 1% of its total data. As Figure 7 shows the multiscale convolution proposed by this paper is better in feature mining and convergence.

E. EXPERIMENTAL RESULTS OF FEATURE INTEGRATION OF DIFFERENT MODULES

In this section, we compare a variety of sequence models. Table 8 shows that when using the sequence model, the experimental effect increases significantly, especially in the Salinas data set; the experimental results are very clear. As the table clearly shows, the experimental results obtained by integrating the GRU module are better than the results obtained by integrating the Bi-GRU module; however, the results obtained by integrating Bi-LSTM are better than the results obtained by integrating LSTM. We suspect that the cell state in LSTM retains more pre-and post-sequence data. The Bi-LSTM we use is more suitable for building feature sequences.

F. EXPERIMENTAL RESULTS OF DIFFERENT SUBMODULES

Table 9 contains the experimental results of different submodels of the three data sets(Indian pine data set: 15% of the total sample number is used as the training set; University of Pavia data set: 10% of the total sample number is used as the training set; Salinas data set: 10% of the total sample number is used as the training set). From the three data sets we can see that all the sub-modules play an active role. In the Indian dataset, the feature dimension after dimensionality reduction is high, and it is very important to construct the context relationship between the features, so the BILSTM algorithm has a greater impact; The effects of each sub-module in the University of Pavia data set are similar, and three sub-modules work best when used together; In Salinas data set, since the reduced dimension has fewer feature dimensions, the integrated attention mechanism becomes more important. From the experimental results, it can be seen that when the integrated attention mechanism is removed, the recognition effect decreases severely. The results are shown in the following figures for different data sets (Figure 8 for Indian Pine dataset; Figure 9 for Pavia University dataset and Figure 10 for the Salinas dataset).

V. CONCLUSION

In this paper, we propose a multitask AML algorithm model (The Deep features of Multi-scale convolution under Attention Mechanism are integrated by Bi-LSTM, AML) that performs feature selection, feature mining and feature integration. The AML algorithm uses an integrated attention mechanism to retain key features while reducing many redundant features and uses multiscale convolution to retain more deep features. Finally, the multiscale deep features are serialized by Bi-LSTM, which strengthens the correlation between the deep features. The best results were achieved on three public data sets and with three improved algorithms. The results fully prove the feasibility of our method.

ACKNOWLEDGMENT

The authors would like to thank the editor-in-chief, the associate editor, and the reviewers for their insightful comments and suggestions.

REFERENCES

[1] O. Ozdil, Y. E. Esin, B. Demirel, and S. Ozturk, "Representative signature generation for plant detection in hyperspectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2709–2712.

[2] S. Z. Mishu, M. A. Hossain, and B. Ahmed, "Hybrid sub-space detection technique for effective hyperspectral image classification," in *Proc. Int. Conf. Comput., Commun., Chem., Mater. Electron. Eng. (IC4ME2)*, Feb. 2018, pp. 1–4.

[3] X. Liu, C. Wang, Q. Sun, and M. Fu, "Target detection of hyperspectral image based on convolutional neural networks," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9255–9260.

[4] C. Zhang, M. Han, and M. Xu, "Multi-feature classification of hyperspectral image via probabilistic SVM and guided filter," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.

[5] Y. Wang, J. Mei, L. Zhang, B. Zhang, P. Zhu, Y. Li, and X. Li, "Self-supervised feature learning with CRF embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2628–2642, May 2019.

[6] M. Salman and S. E. Yuksel, "Fusion of hyperspectral image and LiDAR data and classification using deep convolutional neural networks," in *Proc. 26th Signal Process. Commun. Appl. Conf. (SIU)*, May 2018, pp. 1–4.

[7] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[8] G. Yang, U. B. Gewali, E. Ientilucci, M. Gartley, and S. T. Monteiro, "Dual-channel densenet for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 2595–2598.

[9] F. Nie, S. Yang, R. Zhang, and X. Li, "A general framework for auto-weighted feature selection via global redundancy minimization," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2428–2438, May 2019.

[10] K. Chen and P. Fu, "The feature selection in rolling bearing fault diagnosing based on parts-principle component analysis," in *Proc. 5th Int. Conf. Natural Comput.*, 2009, pp. 613–616.

[11] M. F. I. Ibrahim and A. A. Al-Jumaily, "PCA indexing based feature learning and feature selection," in *Proc. 8th Cairo Int. Biomed. Eng. Conf. (CIBEC)*, Dec. 2016, pp. 68–71.

[12] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, 2018.

[13] M. Z. F. Nasution, O. S. Sitompul, and M. Ramli, "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification," *J. Phys., Conf.*, vol. 978, Mar. 2018, Art. no. 012058.

[14] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.

[15] C. Zhang, J. Wang, Y. Zhang, and Y. Liu, "Small-sample classification of hyperspectral data in a graph-based semi-supervision framwork," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3194–3197.

[16] M. S. Aydemir and G. Bilgin, "Semisupervised hyperspectral image classification using small sample sizes," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 621–625, Feb. 2017.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[18] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[19] Z. Jiang, W. D. Pan, and H. Shen, "LSTM based adaptive filtering for reduced prediction errors of hyperspectral images," in *Proc. 6th IEEE Int. Conf. Wireless for Space Extreme Environments (WiSEE)*, Dec. 2018, pp. 158–162.

[20] Z. Yuan, S. Wu, F. Wu, J. Liu, and Y. Huang, "Domain attention model for multi-domain sentiment classification," *Knowl.-Based Syst.*, vol. 155, pp. 1–10, Sep. 2018.

[21] Y. Qin, G.-W. Shen, W.-B. Zhao, Y.-P. Chen, M. Yu, and X. Jin, "A network security entity recognition method based on feature template and CNN-BiLSTM-CRF," *Frontiers Inf. Technol. Electron. Eng.*, vol. 20, no. 6, pp. 872–884, Jun. 2019.

[22] J. Feng, D. Li, J. Chen, X. Zhang, X. Tang, and X. Wu, "Hyperspectral band selection based on ternary weight convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 3804–3807.

[23] Y. Zhan, H. Tian, W. Liu, Z. Yang, K. Wu, G. Wang, P. Chen, and X. Yu, "A new hyperspectral band selection approach based on convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3660–3663.

[24] J. Feng, X. Wu, J. Chen, X. Zhang, X. Tang, and D. Li, "Joint multilayer spatial-spectral classification of hyperspectral images based on CNN and convlstm," in *Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 588–591.

[25] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.

[26] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[27] H. You, S. Tian, L. Yu, and Y. Lv, "Pixel-level remote sensing image recognition based on bidirectional word vectors," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1281–1293, Feb. 2020.

[28] J. Chen, S. Chen, P. Zhou, and Y. Qian, "Deep neural network based hyperspectral pixel classification with factorized spectral-spatial feature representation," *IEEE Access*, vol. 7, pp. 81407–81418, 2019.

[29] H. Gao, Y. Yang, D. Yao, and C. Li, "Hyperspectral image classification with pre-activation residual attention network," *IEEE Access*, vol. 7, pp. 176587–176599, 2019.

[30] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, "Simultaneous spectral-spatial feature selection and extraction for hyperspectral images," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018.

[31] C. Ji, M. Ye, H. Lu, F. Yao, and Y. Qian, "Feature extraction of hyperspectral imagery based on deep NMF," in *Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 1092–1095.

ZESONG WANG was born in Qingdao, Shandong, China, in 1981. She received the bachelor's degree from Tiangong University, in 2004, and the master's degree from the Shandong University of Science and Technology, in 2010. She has been engaged in Research and Teaching in big data with the College of Qingdao, Huanghai University. Her areas of interests include deep learning and remote sensing image segmentation.

CUI ZOU was born in Qingdao, Shandong, China, in 1983. She received the master's degree in software engineering from the Qingdao University of Science and Technology, in 2015. She currently teaches with Qingdao Huanghai University. Her research interests include deep learning and remote sensing image segmentation.

WEIWEI CAI (Graduate Student Member, IEEE) is currently pursuing the master's degree with the Central South University of Forestry and Technology, Changsha, China. Prior to that, he worked in IT industry for more than ten years in the roles of an Enterprise Architect and a Program Manager. His research interests include machine learning, deep learning, and computer vision.