IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 5, 2020, accepted August 13, 2020, date of publication August 18, 2020, date of current version September 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017641

Automatic Modulation Classification Scheme Based on LSTM With Random Erasing and Attention Mechanism

YUFAN CHEN[1], WEI SHAO[1], JIN LIU[2], (Graduate Student Member, IEEE), LU YU[1], AND ZUPING QIAN[1], (Member, IEEE)

[1]College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China
[2]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

Corresponding author: Wei Shao (tytglpp@126.com)

This work was supported in part by the Jiangsu Provincial Natural Science Foundation of China under Grant BK20160080.

ABSTRACT Automatic modulation classification (AMC) is a key technology of cognitive radio used in non-cooperative communication. Recently, deep learning has been applied to AMC tasks. In this paper, an AMC scheme based on deep learning is proposed, which combines random erasing and attention mechanism to achieve high classification accuracy. Firstly, we propose two data augmentation methods, random erasing at sample level and random erasing at amplitude/phase (AP) channel level. The former replaces training samples with noise information, while the latter replaces AP channel information of training samples with noise information. Erased data segments are randomly stitched to enable training data expansion. Training data of different qualities enables deep learning model to have stronger generalization capability and higher robustness. Then, we propose a single-layer Long Short-Term Memory (LSTM) model based on attention mechanism. In the first part of this model, we propose the signal embedding, which enables the input to contain modulation information more comprehensively and accurately. Then hidden state output by LSTM is input into the attention module, and weighting is applied to the hidden state to help the LSTM model capture the temporal features of modulated signals. Compared to a model without attention mechanism, this model has faster convergence speed and better classification performance. Lastly, we propose a random erasing-based test time augmentation (RE-TTA) method. Test data is randomly erased for multiple times and classification results are comprehensively evaluated, in order to further improve classification accuracy. Experimental results on dataset RML2016.10a show that classification accuracy of the proposed scheme is competitive compared with the state-of-the-art methods.

INDEX TERMS Automatic modulation classification, random erasing, long short-term memory, attention mechanism.

I. INTRODUCTION

As wireless communication technology continues to evolve, electromagnetic environment becomes increasingly complex and volatile. The data amount of modulated signal increases drastically, and the types of modulation mode become increasingly complicated and diversified. Traditional modulation classification algorithms focusing on manual extraction of expert features cannot be applied to the largely emerging communication data, and their classification accuracy is not as good as expected. Therefore, a high-precision data-driven modulation classification algorithm can play a critical role in both civil and military communication. In recent years, some researchers have started to address challenges in wireless communication with deep learning methods, and have achieved certain results. Deep learning based automatic modulation classification (AMC) algorithm not only overcomes the shortcoming that traditional methods cannot use a large amount of communication data, but also enables communication systems to recognize complicated and diversified modulation modes with higher accuracy [1], [2].

In the study of deep learning based AMC technology, in-phase and quadrature phase (IQ) data of modulated signal is used to train deep learning models, and certain results

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

are achieved. West and O'shea [3] used pure IQ data directly for simulation experiments, demonstrated the performance of convolutional neural network (CNN), Residual Network (ResNet) and other network structures in modulation classification tasks, and put forth that classification accuracy is not limited by network depth for the first time. Using IQ data alone for modulation classification poses a certain limitation to the improvement of classification accuracy. With the development of deep learning based AMC technology, researchers have started to improve classification performance in two aspects: data preprocessing and deep learning model improving.

A. DATA PREPROCESSING

1) TRADITIONAL METHODS

Some researchers combine manual feature extraction in traditional methods with deep learning models, and extract expert features from IQ data, hence to classify modulated signals. Peng *et al.* [4] proposed to represent modulated signals with binary constellation diagrams, and optimize constellation representation with gray image and three-channel image. Wang *et al.* [5] used constellation diagrams to classify modulation modes that are difficult to distinguish, such as QAM16 and QAM64. They also proposed to capture data distribution differences of these two modulation modes with a density window. High-order cyclic spectra of modulated signal are also one of the common expert features. Wu *et al.* [6] extracted constellation diagrams and cyclic spectra of modulated signal respectively, and built a two-branch CNN on these two features. Rajendran *et al.* [7] proposed that a good accuracy could be achieved by converting IQ data into amplitude/phase (AP) information and using a simple Long Short-Term Memory (LSTM) model. The model was able to obtain temporal features of signals from training data without requiring manual extraction of expert features.

2) DATA AUGMENTATION METHODS

In the field of deep learning, if model complexity is relatively high and the amount of training data is not sufficient, the risk of overfitting exists. Data augmentation methods can be adopted to expand training data. In addition, data augmentation technology also contributes to classifying data that is easily confusing, and improves model generalization capabilities. In image recognition, audio recognition and other classification tasks, researchers proposed several data augmentation methods, such as flipping [8], random cropping [9], and dropout [10]. Zhong *et al.* [11] proposed an attractive data augmentation method, random erasing, which improved model robustness without introducing any extra learning parameters. Compared to random cropping, random erasing does not impact data integrity. In the field of AMC, Huang *et al.* [12] effectively expanded training data by rotating constellation diagram and adding noise to it. Chen *et al.* [13] used generative adversarial network (GAN) [14] for data augmentation. They proposed an auxiliary classification GAN (ACGAN) in combination with an auxiliary weighting loss function to balance the impact of generated data on classification model, and achieved a classification accuracy of 94% on the open source dataset RML2016.10a [15].

B. IMPROVEMENT OF DEEP LEARNING MODEL

Nie *et al.* [16] proposed a deep hierarchical network (DHN), which combined shallow features with high-level features. It could simultaneously have global receptive field and location information through multi-level feature extraction without any transformation of the raw data. They also innovatively proposed a loss function using signal-to-noise ratio (SNR) as a weight. Simulation experiments showed that the DHN model had relatively high classification accuracy.

As an important concept in the field of deep learning, attention mechanism (AM) was originally used in machine translation [17], and is being widely adopted now in natural language processing (NLP), statistical learning, speech recognition, computers and other fields. In [13], Chen *et al.* proposed a novel attention cooperative framework, which added a self-attention mechanism and a Squeeze-and-Excitation (SE) block [18] to classifiers to obtain interdependencies between input feature maps, and demonstrated effectiveness of attention mechanism in AMC.

C. CONTRIBUTION OF THE PROPOSED AMC SCHEME

In this paper, we propose a deep learning based AMC scheme. As the first step, we extract AP information of IQ data, and then expand training data with two random erasing algorithms. Following that, we obtain classification results through a single-layer LSTM model based on the attention mechanism. Lastly, we further improve classification accuracy with a random erasing-based test time augmentation (RE-TTA) method. The major contributions of our scheme are as follows:

- Random erasing algorithms at sample level and AP channel level for AMC are proposed. They are different from the existing work in [11]. The proposed algorithms enable the proposed scheme to adapt to the increasingly complex environment. Because during the transmission of radio signals, fading, multi-path, noise and other error effects may appear due to irregular terrain and building obstacle, thereby reducing the quality of data collected. In case when all the training data are clearly visible, i.e., no occlusion happens, the learned model may have a good classification accuracy on the data without occlusion, but due to the limited generalization ability of the model, it may not be able to recognize data which are partially occluded. Although we can manually collect signal data in real electromagnetic environments, it is expensive and the level of occlusion might be limited. To summarize, the two random erasing algorithms proposed in this paper have four advantages:

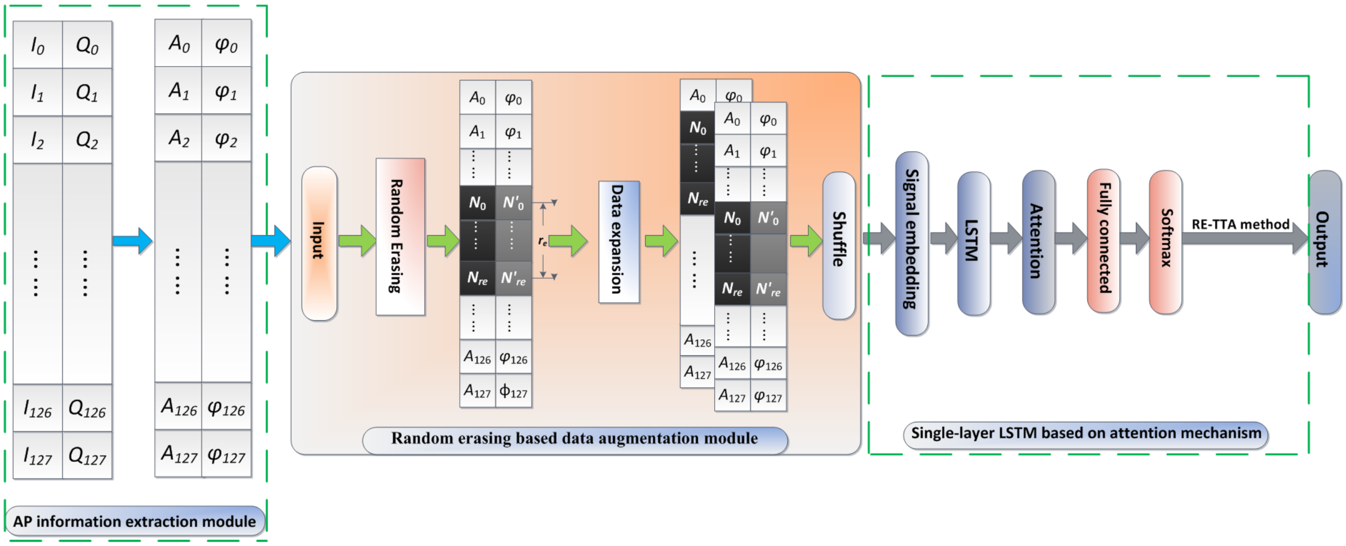

FIGURE 1. System model.

a. The proposed algorithms are specifically designed for AMC and are better suited for processing long sequence signal data.

b. By generating training data of different qualities, the proposed algorithms can significantly improve robustness of deep learning model in front of various error effects in real electromagnetic environments.

c. The proposed algorithms can generate signal data with different occlusion level. Through the training of different quality data, our model has better generalization ability than models trained with high-quality and clean data.

d. The proposed algorithms do not require any extra parameter learning. They can be conveniently embedded into any deep learning model without changing the structure of the model.

- Single-layer LSTM based on attention mechanism is proposed. In deep learning based multi-classification tasks, training of long-sequence data will result in reduction of model classification accuracy, and improper input sequence of training data will also lead to low computational efficiency. In response to such issues, attention mechanism is added to enable deep learning model to capture temporal features of long-sequence data in a way faster than the convergence speed of traditional single-layer LSTM, greatly saving time spent on training. Then we propose signal embedding for AMC. Embedding enables the input vector to include the modulation information of signals more comprehensively and accurately, so as to facilitate the extraction of temporal features by LSTM.
- A RE-TTA method is proposed. In the testing phase, we use random erasing algorithms to generate test data of different qualities and make comprehensive judgments on the corresponding classification results, in order to obtain the final classification results. Classification accuracy is further improved with this method.

The paper is organized as follows: In Section II, the system model is introduced. In Section III, the proposed AMC scheme is described in detail. The simulation results and analyses are shown in Section IV. Finally, some conclusions are presented in Section V.

II. SYSTEM MODEL

In this paper, the system model consists of three parts, as shown in Figure. 1. The first part is the AP information extraction module, the second part is the data augmentation module, and the third part is the classification module. The first two parts are mainly for data preprocessing, and the third part is for realization of modulation classification. As the first part of system model, the AP information extraction module extracts AP information of signals from the IQ data.

As the second part of system model, a data augmentation module is constructed based on random erasing algorithm at sample level or AP channel level. For each batch of input data, a random number is firstly generated in the range of 0 to 1. If the random number is greater than the erasing threshold, the data will be randomly erased, otherwise the data remains unprocessed. After that, data expansion is realized through the random erasing algorithm. To prevent the same batch of data from being repetitively input into the subsequent deep learning model, we perform a Shuffle operation on the expanded data to scramble the data, hence ensuring generalization capability of the model.

As the third part of system model, the classification module is composed of the attention mechanism based LSTM submodule and the RE-TTA submodule. Signal embedding multiply original IQ data and a learnable matrix, so that the data input into LSTM contain more comprehensive modulation information. LSTM is suitable for long-sequence data processing, and can extract temporal features in AP information simultaneously. Among the hidden states output by LSTM, some of the hidden states have extracted the modulation information and are hence activated, while the others remain deactivated. The addition of attention mechanism enables the activated hidden states to get high weight through repeated training, hence getting more attention from the deep learning model. The RE-TTA method is used after the Softmax layer, which generates multiple versions of test set through random erasing, and comprehensively evaluates these classification results to obtain the final result, in order to further improve classification accuracy.

III. THE PROPOSED AMC SCHEME

A. AP INFORMATION EXTRACTION

Assuming the length of a set of data samples is N, firstly, we convert IQ signals into AP information, where amplitude is:

$$A_i = \sqrt{I_i^2 + Q_i^2}, \tag{1}$$

where I_i and Q_i represent the i^{th} data in the sample and A_i the amplitude of the i^{th} data. Then, perform L2 normalization to the data. L2 norm of the amplitude of the i^{th} data can be expressed as:

$$A_{norm} = \sqrt{A_1^2 + A_2^2 + \cdots A_N^2}. \tag{2}$$

Amplitude after L2 normalization is A':

$$A' = A_i / A_{norm}. \tag{3}$$

Phase is expressed as:

$$\varphi_i = \tan^{-1}(Q_i/I_i). \tag{4}$$

The range of phases is -pi to pi, and after normalization of L2 norm, the phase range is -1 to 1

B. RANDOM ERASING BASED DATA AUGMENTATION

The data in RML2016.10a have no occlusion. However, in the actual signal acquisition process, the signal we get will be partially occluded. There are many factors contributing to this phenomenon, such as terrain and other electromagnetic signal interference. Most of the existing classification models do not consider this problem. But a strong classification model should be able to recognize categories from the overall object structure. In order to solve the problem of data occlusion in the actual electromagnetic environment and improve the generalization ability of our trained model, we propose two novel random erasing algorithms, sample level erasing and AP channel level erasing as shown in Figure. 2. These two algorithms mainly simulate different environmental factors. Sample level erasing simulates noise caused by irregular terrain. Noise information is block noise, which is directly generated to replace amplitude and phase information. The AP channel level erasing simulates noise generated by the electromagnetic environment. Random erasing is performed at a certain probability, where the probability of data being randomly erased is p, and the probability of data remaining unchanged is $1-p$. Meanwhile, hyper-parameters r_1, r_2 are defined to control the erasing region.

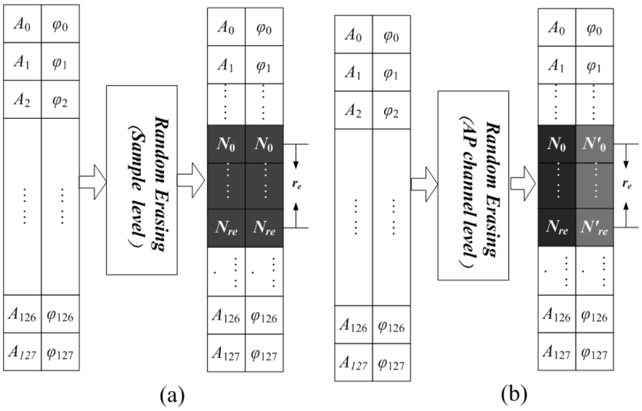

FIGURE 2. Random erasing ((a). at sample level (b). at AP channel level).

Algorithm 1 Sample Level Erasing for AMC

```
Input:
    Input data $\mathbf{D}$;
    Data size $W$ and $H$;
    Erasing rate $P$;
    Erasing aspect ratio range $r_1$ and $r_2$.
Output:
    Erased data $\mathbf{D}'$.
Initialization: $P_0 \leftarrow Rand(0,1)$;
if $P_0 \geq P$ then
    $\mathbf{D}' \leftarrow \mathbf{D}$;
    return $\mathbf{D}'$;
else
    $\mathbf{D}_{min} = Min(\mathbf{D})$;
    $\mathbf{D}_{max} = Max(\mathbf{D})$;
    while True do
        $W_r \leftarrow randint(0, W)$;
        $H_r \leftarrow randint(0, H)$;
        $r_e \leftarrow randint(\frac{H_r}{r_1}, \frac{H_r}{r_2})$;
        if $W_r + r_e \leq w$ then
            $\mathbf{D}_r \leftarrow (W_r, H_r, W_r + r_e, H_r)$;
            $\mathbf{D}(\mathbf{D}_r) \leftarrow Rand(\mathbf{D}_{min}, \mathbf{D}_{max})$;
            $\mathbf{D}' \leftarrow \mathbf{D}$;
            return $\mathbf{D}'$;
        end if
    end while
end if
```

1) RANDOM ERASING AT SAMPLE LEVEL

As shown in Algorithm 1, for random erasing at sample level, only one set of noise points need to be generated in order to replace the AP information to be erased. For a set of data input with a random erasing module, we assume that the data length is n, then amplitude of the signal is $\mathbf{A} = \{A_0, A_1, \cdots\cdots, A_{n-1}, A_n\}$ and phase is $\boldsymbol{\varphi} = \{\varphi_0, \varphi_1, \cdots\cdots \varphi_{n-1}, \varphi_n\}$. The algorithm generates a set of random noise points $\mathbf{N} = \left\{N_0, N_1, \cdots\cdots N_{r_e-1}, N_{r_e}\right\}$, which are used as noise data to replace the AP information of the raw signal:

$$\begin{cases} A_{W_r}, \varphi_{W_r} = N_0 \\ A_{W_r+1}, \varphi_{W_r+1} = N_1 \\ \cdots\cdots \\ A_{W_r+r_e+1}, \varphi_{W_r+r_e+1} = N_{r_{e-1}} \\ A_{W_r+r_e}, \varphi_{W_r+r_e+1} = N_{r_e}. \end{cases} \tag{5}$$

Algorithm 2 AP Channel Level Erasing for AMC

```
Input:
    Input data $\mathbf{D}$;
    Data size $W$ and $H$;
    Erasing rate $P$;
    Erasing aspect ratio range $r_1$ and $r_2$.
Output:
    Erased data $\mathbf{D}'$.
Initialization: $P_0 \leftarrow Rand(0, 1)$;
if $P_0 \geq P$ then
    $\mathbf{D}' \leftarrow \mathbf{D}$;
    return $\mathbf{D}'$;
else
    $\mathbf{D}_{min} = Min(\mathbf{D})$;
    $\mathbf{D}_{max} = Max(\mathbf{D})$;
    while True do
        $W_r \leftarrow randint(0, W)$;
        $H_r \leftarrow randint(0, H)$;
        $r_e \leftarrow randint(\frac{H_r}{r_1}, \frac{H_r}{r_2})$;
        if $W_r + r_e \leq w$ then
            $\mathbf{D}_r \leftarrow (W_r, H_r, W_r + r_e, H_r)$;
            $A(\mathbf{D}_r) \leftarrow Rand(\mathbf{D}_{min}, \mathbf{D}_{max})$;
            $P(\mathbf{D}_r) \leftarrow Rand(\mathbf{D}_{min}, \mathbf{D}_{max})$;
            $\mathbf{D}'(A, P) \leftarrow \mathbf{D}(A(\mathbf{D}_r), P(\mathbf{D}_r))$;
            return $\mathbf{D}'(A, P)$;
        end if
    end while
end if
```

2) RANDOM ERASING AT AP CHANNEL LEVEL

As shown in Algorithm 2, unlike random erasing at the sample level, random erasing at AP channel level does not directly replace each sample, instead, it replaces the AP information of each sample separately and gives a set of noise data $\mathbf{N}' = \left\{N'_0, N'_1, \cdots\cdots \ N'_{r_e-1}, N'_{r_e}\right\}$. The specific random erasing operation is as follows:

$$\begin{cases} A_{W_r} = N_0 \\ A_{W_r+1} = N_1 \\ \cdots\cdots \\ A_{W_r+r_e+1} = N_{r_{e-1}} \\ A_{W_r+r_e} = N_{r_e}, \end{cases} \tag{6}$$

$$\begin{cases} \varphi_{W_r} = N'_0 \\ \varphi_{W_r+1} = N'_1 \\ \cdots\cdots \\ \varphi_{W_r+r_e+1} = N'_{r_{e-1}} \\ \varphi_{W_r+r_e+1} = N'_{r_e}. \end{cases} \tag{7}$$

In our experiment, these noise points are all gaussian white noise with mean of 0 and variance of 1.

3) RANDOM ERASING BASED DATA EXPANSION

In multi-classification tasks, data augmentation technology plays a significant role in improving classification accuracy. Effective data expansion not only increases the quantity of training samples, but also diversifies training samples, effectively minimizing the risk of overfitting. Data expansion methods are designed according to the following principles:

- Improve classification accuracy of deep learning model. Data expansion technology should be capable of improving classification accuracy of deep learning model to a certain level.
- Maintain a lower computational overhead. When training and deploying a deep learning model, we need to consider not only the classification accuracy of the model, but also the space required by training parameters of the model, the memory space required for running the model, the running speed of the model, and *et al.* Data expansion methods should maintain a relatively low computational overhead while improving classification accuracy of the model.

The random erasing algorithms at both sample level and AP channel level proposed in this paper can be used to realize data expansion. Firstly, select the data segment to be expanded, apply random erasing algorithm at sample level or AP channel level to it, perform a Shuffle operation on the expanded data, that is, randomly stitching multiple data segments that have been expanded, and input stitched data as training data for the deep learning model.

C. SINGLE-LAYER LSTM MODEL BASED ON ATTENTION MECHANISM

LSTM is a special type of recurrent neural network (RNN). An LSTM cell features three types of gates, among which the forget gate conditionally decides to discard certain information of the cell, the input gate conditionally decides to update the value of memory state from the input, and the output gate conditionally outputs. By adding these three gating mechanisms, LSTM can effectively capture temporal features of the training data. At the same time, LSTM also solves the problem of vanishing gradient and exploding gradient during the training of long sequences [19]. A single-layer LSTM based on attention mechanism for AMC is proposed as shown in Figure. 3. The first part is signal embedding module. The data format in RML2016.10a is 2 × 128, which can be directly used as input to LSTM, so the existing work such as [7], [12], [13] is to input IQ data directly into LSTM

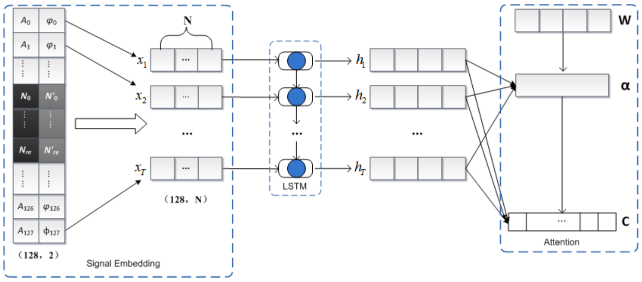

FIGURE 3. The proposed single-layer LSTM based on attention mechanism for AMC.

for processing. If a set of data is viewed as a sentence, then for the LSTM, the sentence is made up of 128 words, each of which is a vector of length 2. Signal embedding is actually a fully-connected operation, which is multiplies the data with a learnable matrix. The reason to use signal embedding is that the features of low-dimensional data are very general. Consequently, we need to continuously increase the data dimensions to enhance the recptive field of the model. The change of data dimension is obtained through continuous learning of the model, and the model will eventually find a dimension that is most suitable for feature extraction. So after embedding, the input matrix will contain more comprehensive and accurate modulation information. The second part is single- layer LSTM. Modulated signals are time series data, and different modulation modes exhibit different amplitude and phase information. LSTM can well extract these temporal features. The last part is the attention mechanism module. For a modulation signal sequence, the amplitude and phase information of part of the data can reflect the feature of the modulation mode. Attention mechanism can pay more attention to this part of data. Suppose we have a sequential signal data which has T points as the input. The first step is to represent the signal as a sequence of T signal embeddings:

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_t}, \ldots, \mathbf{x_T}\}. \tag{8}$$

where $\mathbf{x_t}$ is a N dimensional valued vector, denoting the embedding for the t^{th} signal point in the input. Then the signal data can be represented as a T-by-N matrix, which is the concatenation of the signal embeddings in it. After that, we feed the matrix $\mathbf{X}$ to the LSTM:

$$\mathbf{h_t} = \mathit{LSTM}\left(\mathbf{x_t}, \mathbf{h_{t-1}}\right). \tag{9}$$

Hence the output hidden state of LSTM is

$$\boldsymbol{H} = \{\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \cdots, \mathbf{h}_{T-1}, \mathbf{h}_T\}. \tag{10}$$

Trainable parameter vector is $\mathbf{W_1}$ and bias is b, then vector $\mathbf{H}$ is converted into vector $\mathbf{K}$:

$$\mathbf{K} = \mathbf{W}_1 * \boldsymbol{H} + b. \tag{11}$$

Vector of the normalized weight is $\boldsymbol{\alpha}_\mathbf{n}$:

$$\boldsymbol{\alpha}_\mathbf{n} = \mathit{softmax}\left(\mathbf{W_{2n}} * (\tanh(\mathbf{K}))^T\right), \tag{12}$$

where $\mathbf{W}_{2\text{n}}$ is also a trainable parameter vector, $\boldsymbol{\alpha}_\mathbf{n}$ is a vector of length T, then α_{nt} represents the weight of the t^{th} hidden state output by LSTM:

$$\alpha_{nt} = \frac{\exp\left(\mathbf{W_{2n}} \cdot \mathbf{h_t}\right)}{\sum_{t'=1}^{T} \exp\left(\mathbf{W_{2n}} \cdot \mathbf{h_{t'}}\right)}. \tag{13}$$

All weights add up to 1. After obtaining the weight vector α_{nt} through training, we apply the weights to the hidden states output by LSTM, and the output vector $\boldsymbol{c}_\mathbf{n}$ is:

$$\boldsymbol{c}_\mathbf{n} = \boldsymbol{\alpha}_\mathbf{n}\mathbf{H} = \sum_{\mathbf{t}=1}^{\mathbf{T}} \alpha_{nt} \cdot h_t. \tag{14}$$

In this paper, cross-entropy loss function is:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_{iout}, \tag{15}$$

where N represents the total amount of input data, y_i represents the ground truth label of the i^{th} data, $\hat{y}_{iout}$ represents the predicted label of the i^{th} data provided by the model, and L represents the loss function.

D. RE-TTA

It is assumed that m modulation modes exist in a dataset. In the testing phase, classification results are obtained from the output of the fully connected layer and the Softmax layer in the deep learning model. To further improve classification accuracy, we perform N random erasing operations on the test data and test them separately. The predicted results of the i^{th} data provided by the model are, respectively:

$$\left\{\begin{array}{l} \mathbf{P^1} = \left\{p_0^1, p_1^1, p_2^1 \cdots p_{m-1}^1, p_m^1\right\} \\ \mathbf{P^2} = \left\{p_0^2, p_1^2, p_2^2 \cdots p_{m-1}^2, p_m^2\right\} \\ \qquad\quad \cdots\cdots \\ \mathbf{P^n} = \left\{p_0^n, p_1^n, p_2^n \cdots p_{m-1}^n, p_m^n\right\} \\ \qquad\quad \cdots\cdots \\ \mathbf{P^N} = \left\{p_0^N, p_1^N, p_2^N \cdots p_{m-1}^N, p_m^N\right\}. \end{array}\right. \tag{16}$$

The model makes the finally judgement of the i^{th} data as:

$$\hat{y}_i = \underset{j\in\{0,1\cdots,m\}}{\text{argmax}} \sum_{n=0}^{N} P_j^n, \tag{17}$$

where P_i^n represents the probability that the deep learning model predicts the i^{th} data to be a class-j data in the n^{th} prediction. Test time augmentation scheme utilizes a small amount of test time computational overhead, which improves prediction accuracy of the deep learning model.

IV. SIMULATION EXPERIMENTS AND ANALYSIS

A. DATA SOURCE

In order to verify the performance of the scheme proposed in this paper, we have made simulation experiments using the open source radio dataset RML2016.10a. The dataset contains 11 modulation modes, namely 8PSK, AM-DSB, AM-SSB, BPSK, CFPSK, GFSK, PAM4, QAM16, QAM64, QPSK, and WBFM. The signal data is sampled as IQ data, in a total of 220,000 groups. SNR of the dataset lies in the range from -20 dB to 18 dB, with an interval of 2 dB. For each group of modulated signal, the data format is 2 × 128. During data acquisition, a number of error effects are added in channel environment, such as center frequency offset, sample rate offset, additive white Gaussian noise, and multipath fading.

B. SIMULATION EXPERIMENTS

During the experiments, we divide the dataset into a training set and a test set at a ratio of 7:3. Learning rate in the process of model training is set to 0.001, Adam is selected to be the optimizer, training epoch is set to 70, and batch size is set to 128. The specific parameters involved in each module are given in the analysis below. All experiments are completed with a GPU of RTX 2080Ti and a software environment of python 3.6. The deep learning model is built on Keras with TensorFlow as the backend. Simulation experiments consist of four parts. The first part is simulation experiment for the deep learning model proposed in this paper, that is, the single-layer LSTM based on attention mechanism. The second part is simulation experiment for the random erasing based data augmentation method. The third part is simulation experiment for the RE-TTA method. The fourth part is simulation experiment for classification performance comparison of the proposed AMC scheme with other existing schemes. Meanwhile, we have made relevant analysis of the experimental results.

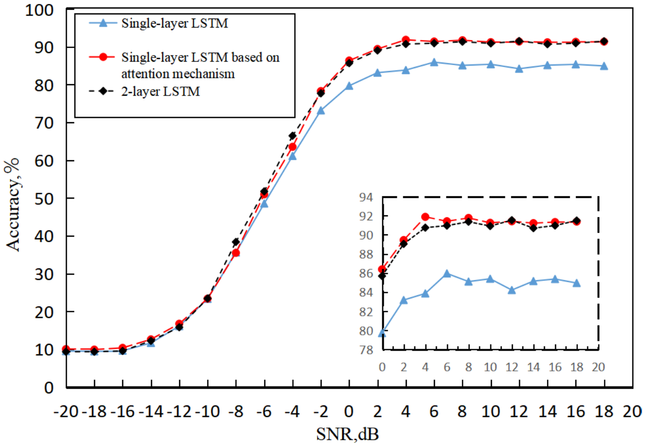

FIGURE 4. Classification accuracy curves of different models.

1) EXPERIMENT OF THE DEEP LEARNING MODEL

As the first step, we convert data into AP information and use it as training data to train the model. During the experiment, performances of three models, namely the single-layer LSTM, the single-layer LSTM based on attention mechanism, and two-layer LSTM, are compared. To prevent overfitting during the experiment, we add a dropout layer with the dropout being set to 0.7. Figure. 4 shows the classification accuracy curves of the three different models. It can be seen that the overall classification accuracy of single-layer LSTM is relatively low. When SNR = 4 dB, its accuracy is 83.96%, while the two-layer LSTM achieves an accuracy of 91.25% when SNR = 4 dB. The single-layer LSTM based on attention mechanism has the highest classification accuracy when SNR = 4 dB, which is 92.08%. According to this, the addition of attention mechanism helps improve classification accuracy of deep learning models. Under the same hardware environment, we have conducted separate experiments on the three deep learning models, and the training parameters, total sample test time are given in Table 1. It can be seen that the single-layer LSTM needs to train 68,491 parameters, and the total testing time of it is 70.288 seconds. The single-layer LSTM model with attention mechanism needs to train 85,131 parameters. Although the addition of attention mechanism increases parameters of the deep learning model, time spent for testing is increased by 0.868 seconds. The two-layer LSTM model needs to train 200,075 parameters, and it takes up to 171.281 seconds to test all the test set, which means that its computational complexity is high. Combining Figure. 4 and Table 1, we find that the single-layer LSTM based on attention mechanism has great advantages in both classification accuracy and computational efficiency. Because the classification accuracy of two-layer LSTM is similar to that of attention based two-layer LSTM, and adding attention mechanism will bring more learned parameters. So we do not add the figure of attention based two-layer LSTM.

TABLE 1. Computational complexity of different models.

Model	Learned parameters	Total sample test time (s)
1-layer LSTM	68491	70.288
1-layer LSTM+AM	**85131**	**71.156**
2-layer LSTM	200075	171.281

* 1-layer LSTM+AM is single-layer LSTM based on attention mechanism

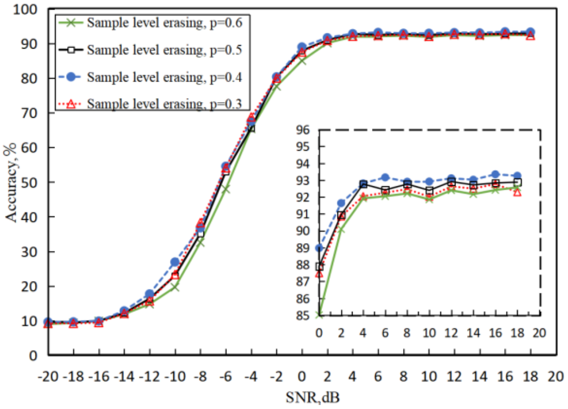

FIGURE 5. Classification accuracy of data augmentation based on random erasing at sample level at different erasing rates.

2) EXPERIMENT OF RANDOM ERASING BASED DATA AUGMENTATION

Erasing rate is an important parameter of the random erasing algorithm, and different erasing rates will have different impact on the training result of deep learning model. Firstly, we determine the value of the optimal erasing rate. Figure. 5

shows classification accuracy curves of the random erasing algorithm at sample level under different erasing rate p. It can be seen that, when erasing rate $p = 0.3$, classification accuracy reaches up to 92.76%. When erasing rate $p = 0.4$, classification accuracy reaches the highest at an SNR of 16 dB, which reaches up to 93.14%; when erasing rate is greater than 0.4, classification accuracy of model declines, with the maximum accuracy reaching up to 92.91% when $p = 0.5$ and 92.56% only when $p = 0.6$. Experiments show that too much intervention on the data leads to a reduction in classification accuracy. Therefore, erasing rate is defined to be $p = 0.4$, and the remaining parameters are set to $r_1 = 4$, $r_2 = 0.25$.

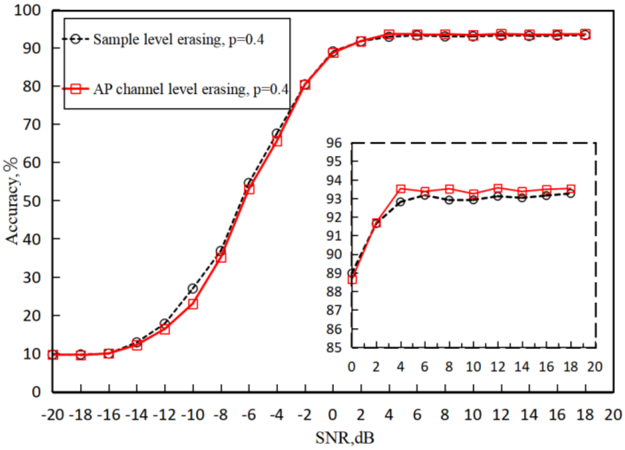

FIGURE 6. Classification accuracy (at $p = 0.4$) of data augmentation schemes based on random erasing at sample level and random erasing at AP channel level.

We use two different random erasing algorithms, namely random erasing at sample level and random erasing at AP channel level to enable data expansion. Erasing rate p is set to 0.4. Figure. 6 shows the classification accuracy comparison of deep learning model using two random erasing algorithms. It can be seen that when SNR $<$ 2 dB, data augmentation based on random erasing at sample level has a higher classification accuracy; when SNR $\geq$ 2 dB, data augmentation scheme based on random erasing at AP channel level has a better classification accuracy, with the average value in the range from 2 dB to 18 dB reaching 93.17% and the highest classification accuracy reaching 93.56%.

Figure. 7 (a) is the confusion matrix obtained by the deep learning model before data is expanded. It can be seen that the deep learning model has a lower classification accuracy for AM-SSB, QAM16, QAM64, WBFM, and AM-DSB modulation modes, which also leads to a low overall recognition rate. We use data expansion to improve classification accuracy of these modulation modes.

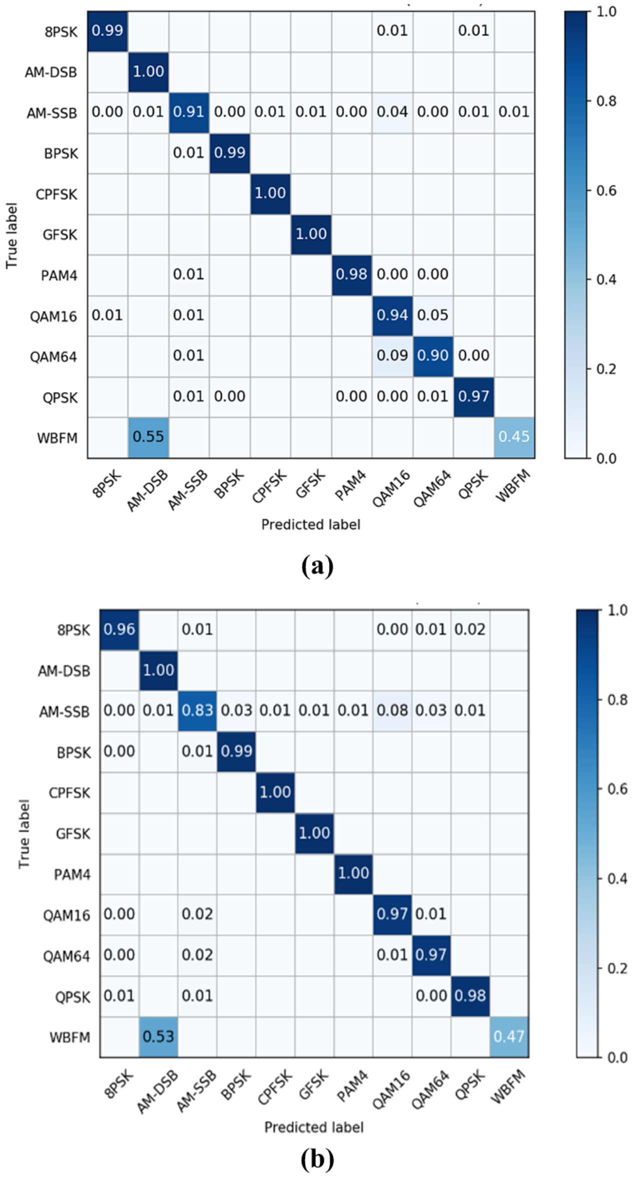

FIGURE 7. Confusion matrixes(at SNR=8dB) ((a). without data expansion (b). with data expansion).

Firstly, we use the random erasing algorithm to expand the data of all modulation modes, and obtain a classification confusion matrix as shown in Figure. 7 (b). It shows that confusion between the two modulation modes (QAM16 and QAM64) is well solved. However, even with an SNR of 8 dB, a small part of data under the modulation mode of AM-SSB will be mistakenly recognized as another modulation mode. Constellation diagram intuitively reflects the distribution of training data. As shown in Figure. 8, we have plotted a constellation diagram for each modulation mode at different SNRs. It can be seen that, data distribution of the various modulation modes at low SNR is difficult to distinguish. In order to further study and verify the above, we have made a simulation and obtained the confusion matrix at low SNR, as shown in Figure. 9. It can be seen that, at low SNR, there is a high probability that the various modulation modes are recognized as AM-SSB.

In the RML2016.10a dataset, modulated signal is generated from real audio stream, therefore, when voice signal is idle, AM-DSB and WBFM tend to be confusing as they do not include useful modulation information [20]. To this end,

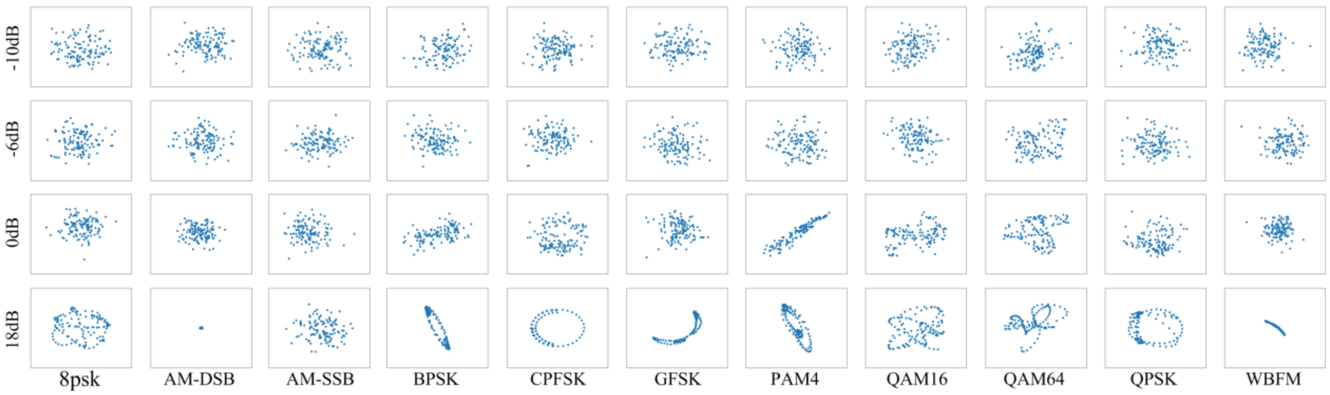

FIGURE 8. Constellation diagrams of eleven modulation modes at different SNRs.

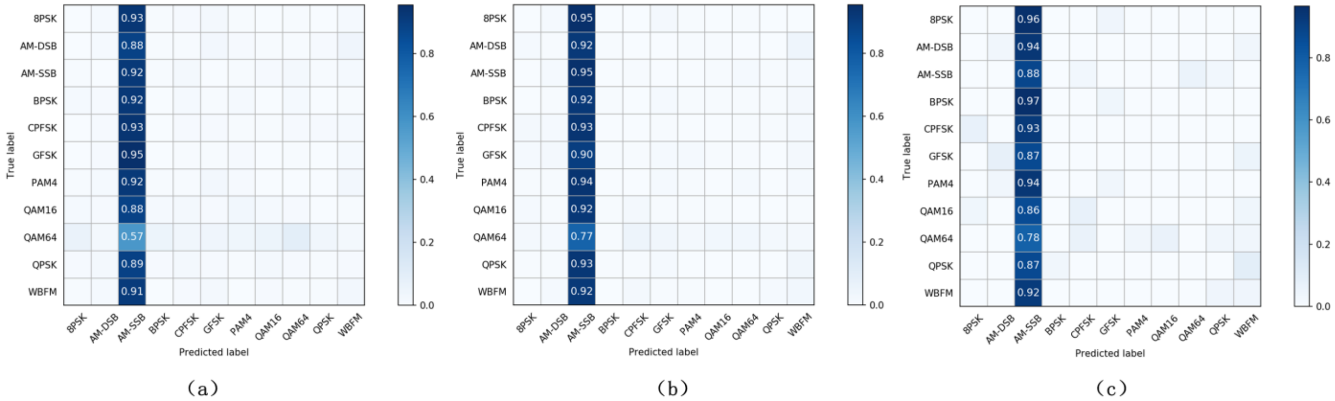

FIGURE 9. Confusion matrixes ((a) SNR = −16 dB, (b) SNR = −18 dB, (c) SNR = −20 dB).

we expand data for AM-SSB, QAM16 and QAM64 in the simulation experiments described below.

TABLE 2. Impact of hyper-parameter N value selection on classification accuracy.

Method	Accuracy	Improvement
Single-layer LSTM	86.14%	–
Attention	92.08%	5.94%
RE based data augmentation	93.56%	1.48%
RE-TTA	94.09%	0.53%

3) EXPERIMENT OF RE-TTA

RE-TTA performs N random erasing operations on the test data to generate N versions of test data. We choose hyper-parameter N through the analysis of experimental results. Table 2 shows the impact of the proposed scheme on classification accuracy when different value is taken for N. When $N = 2$, the maximum classification accuracy of the proposed scheme is 93.852%, which is 0.502% higher than that without RE-TTA. When $N = 4$, the maximum classification accuracy is 93.92%, which is 0.861% higher than that without RE-TTA. When $N = 3$, the maximum classification accuracy is 94.091%, which is 0.690% higher than that without RE-TTA and is the optimal classification accuracy.

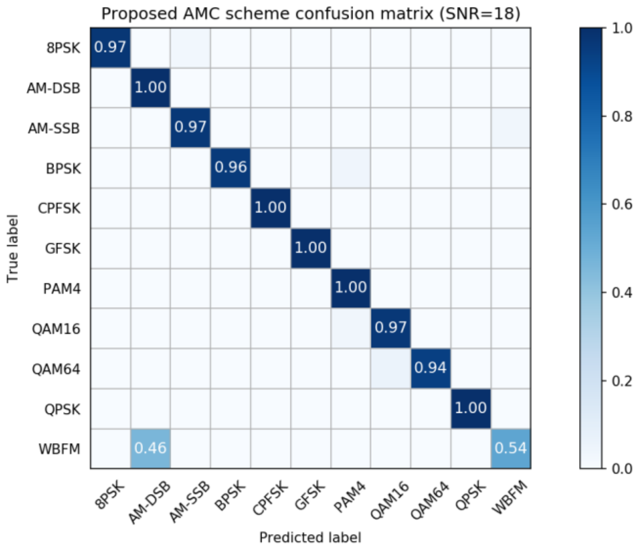

FIGURE 10. Confusion matrix of the proposed AMC scheme at SNR = 18 dB.

4) EXPERIMENT OF THE PROPOSED AMC SCHEME

Figure. 10 shows the confusion matrix of the proposed scheme at SNR = 18 dB. It can be seen that classification accuracy of the proposed scheme reaches 94.091% at SNR = 18 dB. At the same time, it also effectively solves the problem of confusing QAM16 and QAM64, among which, recognition accuracy is 97% for QAM16 and 94% for QAM64. In addition, recognition accuracy of the proposed scheme reaches 97% for AM-SSB, significantly solving the problem of low classification accuracy for the modulation mode. However, confusion still exists with AM-DSB and WBFM.

TABLE 3. Methods used in this paper and their improvements.

Hyper-parameter section	Accuracy improvement under different SNR(%)			
	SNR=6dB	SNR=10dB	SNR=14dB	SNR=18dB
Improvement(N=2)	**0.502(93.852)**	0.41	-0.152	0.57
Improvement(N=3)	0.329	0.434	0.17	**0.861(93.92)**
Improvement(N=4)	0.31	0.472	0.23	**0.690(94.091)**

a: ANALYSIS ON CLASSIFICATION PERFORMANCE OF THE PROPOSED AMC SCHEME

Table 3 lists the contributions of various methods in the proposed scheme in improving classification accuracy, which include:

- Adding an attention mechanism to the single-layer LSTM has the most significant improvement on classification accuracy, reaching 5.943%, and the introduction of attention mechanism makes the model converge more quickly at lower computational complexity.
- Random erasing based data augmentation has brought a classification accuracy improvement of 1.479%, and the introduction of this scheme has well solved the problem of recognizing confusing modulation modes.
- RE-TTA has brought an improvement of 0.531% on model performance.

Due to the above contributions made by the various technologies, maximum classification accuracy achieved by the AMC scheme proposed in this paper has reached 94.091%.

b: CLASSIFICATION PERFORMANCE COMPARISON OF THE PROPOSED AMC SCHEME AND OTHER EXISTING SCHEMES

We have compared the proposed scheme to seven existing schemes, namely CNN-IQ, CNN-AP, ResNet-IQ, LSTM-AP, LSTM-IQ and CLDNN-IQ, CLDNN-AP. Among them, CNN [21] and ResNet [22] can improve the affect of signal frequency offset on classification accuracy; LSTM is suitable for feature extraction of time series signals [19]; while Convolutional, LSTM, Deep Neural Network (CLDNN) combines the advantages of DNN, CNN and LSTM, and is proved to have good performance in the classification of modulation modes [23].

Figure. 11 compares classification accuracy of the proposed scheme to that of other seven schemes: CNN-IQ, CNN-AP, ResNet-IQ, LSTM-AP, LSTM-IQ and CLDNN-IQ, CLDNN-AP. CNN-IQ has a relatively low classification accuracy, its maximum accuracy is 81% only, which shows that CNN is relatively low-performing in feature extraction of time series signals. Even if the amplitude and phase information of IQ signal is extracted as the training data of CNN, no significant effect is achieved, with the highest accuracy of CNN-AP only reaching 83.4%. ResNet-IQ has a slightly higher classification accuracy than CNN-IQ, and its maximum accuracy is 83.5%. The CLDNN-IQ model has a higher classification accuracy than other models at low SNR. When SNR = 0 dB, its classification accuracy is 80.7%. As SNR goes up, its maximum accuracy reaches 85.8%. However, when the amplitude and phase information is used as CLDNN input, the classification accuracy can only reach 85.2%.

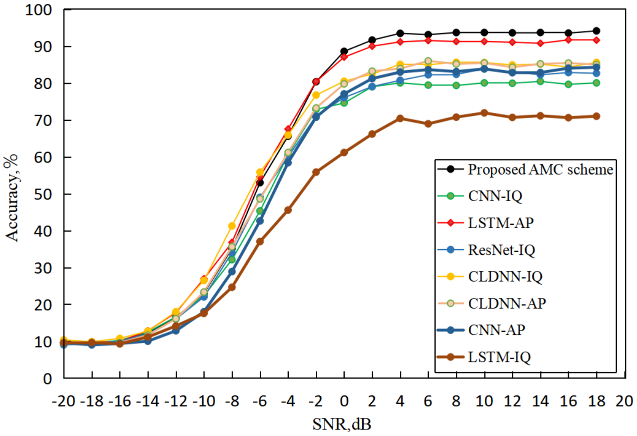

FIGURE 11. Classification performance comparison of the proposed scheme vs. existing schemes.

As can be seen from Figure. 11, the classification accuracy of IQ data as the input of LSTM is low. Because different modulation schemes exhibit different amplitude and phase characteristics, and that's the IQ data does not show. When SNR = 0 dB, classification accuracy is 87.13%; when 0dB $\leq$ SNR $\leq$ 18 dB, average classification accuracy is 90.69%, which is higher than that of the CNN-IQ scheme that uses IQ data for training. The scheme proposed in this paper has achieved an accuracy of 89.2% when SNR = 0 dB, an average classification accuracy of 92.87% when 0dB $\leq$ SNR $\leq$ 18 dB, and a maximum classification accuracy of 94.091%. Simulation results have proved that the proposed scheme is an advanced one in terms of classification accuracy.

V. CONCLUSION

In this paper, a deep learning based AMC scheme is proposed, which combines random erasing and attention mechanism to achieve high classification accuracy. By studying the selection of LSTM layers, a single-layer LSTM model based on attention mechanism is selected on top of the trade-off between computational complexity and classification accuracy. Signal embedding enables our input to contain more accurate and comprehensive modulation information. The addition of attention mechanism contributes greatly to classification accuracy improvement of the deep learning model, and experimental results also demonstrate this model enjoys a significant advantage in temporal feature extraction. To further improve generalization capability and robustness of the model, we have proposed two random erasing based data augmentation schemes. Finally, we have proposed the RE-TTA scheme to further improve classification accuracy of modulation modes. Classification accuracy of the proposed scheme is compared to that of multiple existing schemes on the open source dataset RML2016.10a, and the proposed scheme is proved to be an advanced and effective one.

REFERENCES

[1] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-Air deep learning based radio signal classification," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.

[2] G. J. Mendis, J. Wei, and A. Madanayake, "Deep learning-based automated modulation classification for cognitive radio," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Dec. 2016, pp. 1–6.

[3] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Mar. 2017, pp. 1–6.

[4] S. Peng, H. Jiang, H. Wang, H. Alwageed, Y. Zhou, M. M. Sebdani, and Y.-D. Yao, "Modulation classification based on signal constellation diagrams and deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 718–727, Mar. 2019.

[5] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4074–4077, Apr. 2019.

[6] H. Wu, Y. Li, L. Zhou, and J. Meng, "Convolutional neural network and multi-feature fusion for automatic modulation classification," *Electron. Lett.*, vol. 55, no. 16, pp. 895–897, Aug. 2019.

[7] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 3, pp. 433–445, Sep. 2018.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, p. 12.

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[11] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896

[12] L. Huang, W. Pan, Y. Zhang, L. Qian, N. Gao, and Y. Wu, "Data augmentation for deep learning-based radio modulation classification," *IEEE Access*, vol. 8, pp. 1498–1506, 2020.

[13] S. Chen, Y. Zhang, Z. He, J. Nie, and W. Zhang, "A novel attention cooperative framework for automatic modulation recognition," *IEEE Access*, vol. 8, pp. 15673–15686, 2020.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, and S. Ozair, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[15] T. J. O'Shea and N. West, "Radio machine learning dataset generation with gnu radio," in *Proc. GNU Radio Conf.*, 2016, vol. 1, no. 1, p. 12.

[16] J. Nie, Y. Zhang, Z. He, S. Chen, S. Gong, and W. Zhang, "Deep hierarchical network for automatic modulation classification," *IEEE Access*, vol. 7, pp. 94604–94613, 2019.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, May 2015.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 563–575, Dec. 2017.

[21] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 253–256.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: http://arxiv.org/abs/1512.03385

[23] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580–4584.

YUFAN CHEN received the B.S. degree, in 2018. He is currently pursuing the M.Eng. degree with the College of Communication Engineering, Army Engineering University of PLA, Nanjing, China. His current research interests include deep learning and automatic modulation classification.

WEI SHAO received the B.S., M.S., and Ph.D. degrees from the College of Communications Engineering, Army Engineering University of PLA, Nanjing, China, in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the Army Engineering University of PLA. His current research interests include intelligent spectrum management and communication signal processing.

JIN LIU (Graduate Student Member, IEEE) received the B.S. degree, in 2018. She is currently pursuing the M.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China. Her current research interests include photogrammetry and 3D reconstruction in computer vision.

LU YU received the B.S. degree from the PLA Institute of Communications Engineering, in 1996, the M.S. degree from the PLA University of Science and Technology, in 2000, and the Ph.D. degree from Southeast University, in 2007. She is currently with the Institute of Communications Engineering, Army Engineering University of PLA. Her current research interests include machine learning and image understanding.

ZUPING QIAN (Member, IEEE) was born in Haimen, Jiangsu, China, in 1961. He received the B.S. and M.S. degrees in applied mathematics from Hunan University, Changsha, China, in 1982 and 1985, respectively, and the Ph.D. degree in microwave techniques from Southeast University, Nanjing, China, in 2000. From 1985 to 1999, he was with the Institute of Communications Engineering, Nanjing, as a Lecturer and later as an Associate Professor. Since 2000, he has been a Professor with the College of Communications Engineering, Army Engineering University of PLA, Nanjing. He has authored several books, such as *Electromagnetic Compatibility, Antenna, and Propagation*. He has authored over 80 international and regional refereed journal articles. His research interests include antenna, metamaterials, computational electromagnetics, array signal processing, and EMI/EMC.