Toward Next-Generation Signal Intelligence: A Hybrid Knowledge and Data-Driven Deep Learning Framework for Radio Signal Classification

Shilian Zheng, Xiaoyu Zhou, *Member, IEEE*, Luxin Zhang, Peihan Qi, *Member, IEEE*, Kunfeng Qiu, Jiawei Zhu, and Xiaoniu Yang

***Abstract*—Automatic modulation classification (AMC) can generally be divided into knowledge-based methods and data-driven methods. In this paper, we explore combining the knowledge-based method and data-driven technology to take full advantage of both and propose a hybrid knowledge and data-driven deep learning framework (HKDD) for AMC. To make the handcrafted features more discriminative, various traditional features are adopted, including instantaneous features, statistical features, and spectral features. In the HKDD framework, a feature fusion mechanism is proposed to integrate the features learned from the original signal with those processed by a fully connected network from the handcrafted features. Besides, an attention mechanism is implemented on the fused features to neglect immature features and highlight important features. To evaluate the performance of the proposed method, we construct two modulation classification datasets containing both traditional features and raw IQ data. The bigger one contains 36 modulation categories, which is greater than the number of categories of any AMC dataset currently available. Simulation results show that our proposed method has significant performance gain in both adequate-sample classification scenario and few-shot classification scenario.**

***Index Terms*—Automatic modulation classification, few-shot classification, deep learning, attention mechanism, traditional features.**

I. Introduction

With the explosive development of wireless communication technology, various communication networks have been deployed and intelligent terminals have gradually popularized. Nowadays, millions Internet of Things (IoT) devices have been deployed for providing wireless services, and the number is growing with 25% rate annually, achieving 80 billion by 2030 [1]. The sharp increase of IoT devices has posed a severe challenge to the spectrum resources, that is, it is crucial to accommodate the ever-increasing demand for wireless services and allow a massive amount of IoT devices to access the spectrum. An effective way for alleviating the situation is to use dynamic spectrum access (DSA) [2] technology based on cognitive radio (CR) [3], [4] to improve the spectrum utilization efficiency by allowing the unlicensed user to access the licensed band when licensed user is absent. In the field of DSA, automatic modulation classification (AMC) has become a key technology to optimize spectrum allocation by assisting the unlicensed user to detect the signal of a licensed user without any prior knowledge [5], [6]. When the modulation type of the licensed user is recognized, the unlicensed user can choose an appropriate modulation type for transmission in order to reduce interference to the licensed user. In addition, AMC technology has also been widely used in other fields, including interference identification, communication reconnaissance, and blind signal processing.

Most traditional AMC algorithms are designed based on domain knowledge which may come from presumptive statistical models or deterministic models associated with theory in the field of communications and signal processing. For example, the feature-based AMC methods design handcrafted features based on the deterministic models of the transmitted signal with respect to a specific modulation type while the likelihood-based AMC methods usually rely on the assumed channel model, e.g., additive white Gaussian noise (AWGN) channel [7]. We refer to these methods as knowledge-based methods which mainly rely on the domain knowledge to perform modulation classification. In general, the knowledge-based methods do not rely on a large number of training samples to learn the relationship between the input and the desired output because only a few parameters are required to be estimated. However, the knowledge-based methods are difficult to adapt to the complicated and dynamic channel environment. Besides, they depend too much on selecting applicable features for different modulation types and these features are usually

Manuscript received 25 September 2022; revised 30 December 2022; accepted 5 February 2023. Date of publication 10 February 2023; date of current version 9 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U19B2016, U20B2038 and 62171334. The associate editor coordinating the review of this article and approving it for publication was S. Mao. *(Corresponding authors: Shilian Zheng; Xiaoniu Yang.)*

Shilian Zheng, Luxin Zhang, Kunfeng Qiu, Jiawei Zhu, and Xiaoniu Yang are with the No. 11 Research Center, Science and Technology on Communication Information Security Control Laboratory, Jiaxing 314033, China (e-mail: lianshizheng@126.com; lxzhangMr@126.com; yexijoe@163.com; zhujiaweigl@126.com; yxn2117@126.com).

Xiaoyu Zhou and Peihan Qi are with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an 710071, China (e-mail: zxy0686@126.com; phqi@xidian.edu.cn).

Digital Object Identifier 10.1109/TCCN.2023.3243899

not optimal in recognizing large number of modulation categories.

As the giant success of deep learning (DL) in various applications such as image recognition [8] and text classification [9], it has also been used in radio signal processing including signal detection [10], signal classification [11], [12] and information recovery [13]. The data-driven DL methods for AMC are commonly used in a supervised learning manner. A deep neural network is designed and trained with a mass of labeled samples to extract high-dimensional features from input signals to distinguish different modulation types. The DL networks are generally regarded as a non-linear mapping from the input to the output, and the mapping function specified by a large number of parameters is optimized using the training samples. In general, the data-driven DL methods can usually obtain better performance than the knowledge-based methods when adequate training samples are available. However, note that the DL model is usually high-parameterized, once the number of training samples is insufficient, it will be difficult to find optimal values for the parameters of DL model, leading to a sharp decline of classification performance for the DL-based methods. Hence, the performance of DL-based AMC methods will suffer in a few-shot scenario.

The next generation of artificial intelligence (AI) refers to explainable AI that is able to explain the model behavior and gains insight in the working mechanism by combining domain knowledge [14]. Domain knowledge can provide constructive guidance for adjusting the DL model to improve the related performance. For example, knowledge about statistic properties of raw data was combined with convolutional neural network (CNN) and broad learning to design a fault diagnosis framework in [15]. Domain-specific knowledge of handwritten Chinese characters, including deformation, non-linear normalization, imaginary strokes, and path signature was incorporated with CNN to improve the recognition performance of handwritten Chinese characters in [16]. Knowledge-driven image preprocessing module was introduced for camera recognition in [17] to extract multi-scale knowledge of images. The multi-scale knowledge of these images and the original image are sent to CNN to get the camera type of the picture. These works reveal the potential of combining domain knowledge and data-driven technology to improve the performance of recognition.

In the field of radio signal processing, we envision the next generation of signal intelligence as the hybrid technology that attempts to take the combination of domain knowledge and data-driven technology into account which has not been thoroughly investigated. In this paper, we propose a Hybrid Knowledge and Data-driven Deep learning framework (HKDD) for AMC which can obtain good performance in both adequate-sample scenario where adequate labeled samples are available and few-shot scenario where only a small amount of labeled samples are available. The knowledge we considered in this paper is the handcrafted features explored in feature-based AMC methods, which consist of instantaneous features, statistical features and spectral features. Our proposed network can be divided into three parts: DL network, knowledge network and fusion network. The DL network is similar to the DL-based AMC method, which gives the prediction result based on the IQ signal input. The knowledge network produces the prediction result based on the input of handcrafted features. The fusion network is to combine the learned features from DL network and knowledge network and produce more discriminative joint features. We build two datasets to evaluate the performance of our proposed method. Overall, the contributions of this paper can be summarized as follows.

- In order to promote the classification performance of AMC in both adequate-sample classification scenario and few-shot classification scenario, we propose HKDD to take full advantage of the DL-based method and knowledge-based method through integrating the features learned through a CNN from the original signal with those processed by a deep neural network (DNN) from the handcrafted features.
- To alleviate the influence of some immature features caused by inadequate learning, we adopt an attention mechanism to automatically learn corresponding weights for fused features in our proposed HKDD, which can abandon immature features by learning weights close to 0 and highlight important features by learning weights close to 1.
- We build two datasets for validating our proposed method, namely, HKDD_AMC12 which contains 12 different modulation types and HKDD_AMC36 which contains 36 different modulation types. Besides raw IQ sequences, traditional features of these signals are also included in these datasets. We concatenate traditional features into a vector for the convenience of DNN processing. The number of categories of HKDD_AMC36 is greater than the number of categories of any AMC dataset currently available.
- We evaluate the performance of our proposed method in both adequate-sample scenario and few-shot scenario. Simulation results show that the proposed HKDD is superior to the DL-based method and the knowledge-based method in both scenarios. Furthermore, HKDD also performs far better than an existing "hybrid" AMC method.

The rest of the paper is organized as follows. We discuss the related work in Section II and introduce the system model in Section III. We give basic definitions of traditional features adopted in our proposed method in Section IV. We explain the details of our proposed HKDD framework in Section V. The modulation datasets and simulation results are given in Section VI and finally the conclusion is made in Section VII.

II. Related Work

A. Knowledge-Based AMC Methods

Among the knowledge-based AMC methods, we focus on the feature-based AMC methods since the handcrafted features have low complexity in computation and easy to implement. In general, the feature-based AMC methods usually utilize several signal features to make a decision and the adopted signal features need to be designed carefully for different

modulations. Examples of the features include instantaneous features, statistical features and spectral features.

Information contained in the instantaneous amplitude, instantaneous phase and instantaneous frequency of the received signal is valuable to discriminate the modulation type and many methods have been proposed to extract this information. In [18], [19], [20], the authors employed the standard deviation of the absolute value of the normalized-centered instantaneous amplitude to classify 2ASK and 4ASK and the standard deviation of the absolute value of the normalized centered instantaneous frequency to distinguish between 2FSK and 4FSK. Phase difference was used to identify the PSK order in [21], [22]. Kurtosis of the amplitude was used for PSK and QAM identification in [23].

The most commonly used statistical features for AMC are high-order cumulants and moments. High-order moments were employed as classification features in [24]. In [25], high-order cumulants were introduced as the discriminative features to distinguish between ASK, PSK, and QAM modulations. In [26], a robust AMC algorithm based on fourth-order cumulants was proposed when multipath fading channel is considered and the prior information on the channel state is unknown. Furthermore, the fourth-order cumulants were extended to eighth-order cumulants in [27], and it has been proved that the eighth-order cumulants-based algorithm can achieve much better classification accuracy in distinguishing PSK, FSK and QAM signals under multipath fading channels. High-order cumulants were also used to classify the modulations of multiple-input multiple-output (MIMO) signals in [28], [29]. Recently, the AMC algorithm based on high-order cumulants has been introduced in distributed networks [30].

Spectral features represent features of signals in the frequency domain, which provides another new perspective to distinguish signals. Two key spectral features were introduced in [31] for recognizing analog modulation signals, the maximum value of the spectral power density of the normalized-centered instantaneous amplitude and the signal spectrum symmetry derived from the signal spectrum. The former feature is used to divide various modulation types into two families. One is the modulation type that the signal amplitude carries information, such as M-PAM and M-QAM, and the other contains modulation types that signal amplitude is unchanged, such as FM, M-FSK. The latter feature is effective to measure the symmetry of the spectrum. Furthermore, discrete Fourier transform (DFT) of the phase histogram was used to identify the PSK order in [32]. The DFT of the phase histogram was used to classify various QAM signals by combining knowledge about the distribution of the magnitude in [33].

B. Data-Driven AMC Methods

With the rapid development of DL, many AMC methods based on DL have been proposed. The data-driven DL methods are commonly trained on massive labeled samples, where the original IQ signal is commonly used as the input, and they aim at designing a deep network to extract high-dimensional features from the raw input signals to distinguish different modulation types. With the advent of some excellent CNN models in the task of image classification, such as AlexNet [34], GoogleNet [35], ResNet [36], many works have explored the usage of CNN to complete the task of modulation classification. An AlexNet based feature learning network was proposed in [37]. It was designed to extract deep features using parameter-based transfer learning techniques for promoting multi-level representation capabilities of features and reducing the requirements of sample size. In [38], the authors designed a special architecture of CNN with 34 layers for AMC. Moreover, the training set was enhanced by means of interpolation, extraction, power normalization, and Gaussian noise to improve the robustness of the recognition algorithm. Due to the superior performance of ResNet in image classification, it has been employed in modulation classification recently in [39], [40], [41] and it works well whether it classifies 24 modulation types or high-order modulations, such as 256QAM and 1024QAM. As the complex convolution can extract amplitude and frequency features from the complex-valued signal, a designed complex-ResNet was used in [42] to recognize multiple modulations of signals. To bridge the gap between the wireless signals and DL models, the authors in [43] proposed to transform complex-valued signal waveforms into contour stellar image (CSI), which can be treated as a general image data format.

Considering the communication signal is actually a temporal sequence and correlated in time, recurrent neural networks (RNNs) have been adopted for AMC. RNN is effective to learn the non-linear characteristics of the time sequence due to its memory mechanism. The authors in [44] focused on extracting time-related characteristics of communication signals by RNN rather than spatial-related characteristics by CNN and compared the performance of CNN, RNN, long short-term memory (LSTM), and gated recurrent unit (GRU) network. A robust AMC method based on RNN was proposed in [45], where the channel noise was considered as a mixture of different noises. As the channel noise was found to be time-related data, the RNN-based method was proved to be superior to the method that requires estimating channel and noise iteratively. A LSTM-based classifier was proposed in [46] for extracting time-related relation with signal sequence without estimation of the signal parameters.

Some works try to integrate different networks in order to boost the performance of AMC. The authors in [44] achieved performance gains through incorporating the RNN into the CNN-based method in both the AWGN channel and Rayleigh fading channel. In [45], a classifier composed of two convolutional layers followed by one LSTM layer was proposed for the modulation recognition. The experimental results reveal that this structure can effectively extract the temporal correlation and the classification performance is better than that without LSTM. The authors in [47] proposed a new AMC method that fused the features extracted from one-dimensional convolution, two-dimensional convolution and LSTM. It improves the accuracy of classification, especially for 16QAM and 64QAM.

C. Hybrid AMC Methods

Hybrid AMC methods try to incorporate the knowledge-based method and the DL-based method to improve the performance of AMC. However, only a few works attempted to take the combination of knowledge and data-driven technology into account. In [48], the authors integrated the handcrafted features with the extracted features by CNN from the time-frequency distribution of the received signal for AMC. However, the handcrafted features considered are limited. The DL model was trained on adequate samples and the authors didn't take manner to ensure the classification performance in the few-shot scenario. More importantly, the raw IQ input was not considered in their hybrid structure which may lead to severe performance loss. The authors in [49] focused on the semi-supervised learning scenario, where some handcrafted features, such as high-order cumulants features, entropy features and time-frequency features were combined with unsupervised features extracted by autoencoder as well as the labeled samples to train an annotator to label the unlabeled samples. As a result, adequate pseudo-labeled samples and a few real-labeled samples were applied to train a classifier. It can be seen that as a practical representation of domain knowledge, the traditional handcrafted features are more favorite by authors since their low complexity of computation and easy implementation. Different from that only the adequate-sample classification scenario was considered in the above works, in this paper we consider both the adequate-sample classification scenario and few-shot classification scenario. Furthermore, in our proposed HKDD, we jointly optimize sub-networks for the handcrafted features input and the IQ data input rather than concatenating the handcrafted features with the extracted features from the IQ data directly, thereby avoiding the influence of excessive value of the handcrafted features on the classification layer of DL model.

III. System Model

Considering a discrete-time baseband equivalent model, the relation between the transmitted signal $s_m(n)$ and the received signal $r(n)$ at time instant n can be expressed as

$$r(n) = s_m(n) * h(n)e^{j(2\pi n\Delta f+\theta_0)} + w(n), \tag{1}$$

where * represents the convolution operation, $s_m(n)$ is the modulated signal which is generated from one of M modulations $\{s_1(n), s_2(n), \ldots, s_M(n)\}$, $h(n)$ is the impulse response of the transmitted wireless channel, which is simply an impulse function $\delta(n)$ for ideal channel, $w(n)$ is complex-valued white Gaussian noise with zero mean and variance σ_n^2, Δf is the frequency offset of carrier, θ_0 is a random phase shift due to frequency offset of carrier and phase jitter, $n = 0, 1, \ldots, N-1$, and N denotes the signal length.

Modulation classification is commonly modeled as a classification problem with M categories. The goal of modulation classification is to recognize the modulation type of transmitted signal $s_m(n)$ using the received signal $r(n)$ and maximize the probability $\Pr(s_m(n) \in \mathcal{M}_i | r(n))$, where $\mathcal{M}_i$ represents the i-th modulation scheme. For simplicity in implementation and computation, the received signal is generally represented in $N \times 2$ format, where N is the signal length. The in-phase and quadrature components of $r(n)$, also known as IQ components, are stacked in parallel for the convenience of implementation. The IQ components can be represented by

$$\begin{aligned} I(n) &= \mathrm{real}(r(n)), \\ Q(n) &= \mathrm{imag}(r(n)), \end{aligned} \tag{2}$$

where $I(n)$ and $Q(n)$ correspond to the in-phase and quadrature components of $r(n)$ respectively, $\mathrm{real}(\cdot)$ and $\mathrm{imag}(\cdot)$ represent the real and imaginary parts of the signal.

IV. Adopted Traditional Features

We consider hybrid modulation classification scenario where multiple traditional features which are usually derived from domain knowledge are combined with the IQ samples to improve the modulation classification performance. In this section, we give the basic definitions of traditional features adopted in this paper, which can be divided into three categories, namely, instantaneous features, statistical features, and spectral features.

A. Definitions of Instantaneous Features

For a received signal $r(n), n = 0, 1, \ldots, N-1$, where N is equal to the sampling points, the instantaneous amplitude $a(n)$ of received signal is defined as

$$a(n) = \sqrt{I(n)^2 + Q(n)^2}. \tag{3}$$

Instantaneous amplitude used in the paper is normalized-centered instantaneous amplitude $A(n)$ and the operation of normalization is expressed as follows:

$$A(n) = \frac{a(n)}{\mathrm{E}(a(n))} - 1, \tag{4}$$

where $\mathrm{E}(\cdot)$ is to calculate the mean value.

The instantaneous phase is calculated through the following equation:

$$P(n) = \begin{cases} 0, & Q(n) = 0, I(n) > 0, \\ \arctan(Q(n)/I(n)), & Q(n) > 0, I(n) > 0, \\ \pi - \arctan(Q(n)/I(n)), & Q(n) > 0, I(n) < 0, \\ \pi/2, & Q(n) > 0, I(n) = 0, \\ \pi, & Q(n) = 0, I(n) < 0, \\ \pi + \arctan(Q(n)/I(n)), & Q(n) < 0, I(n) < 0, \\ 3\pi/2 - \arctan(Q(n)/I(n)), & Q(n) < 0, I(n) > 0. \end{cases} \tag{5}$$

The instantaneous frequency $f(n)$ is obtained by the difference of instantaneous phase $P(n)$ as

$$f(n) = P(n) - P(n-1), n = 1, 2, \ldots, N-1. \tag{6}$$

In order to keep the length of $f(n)$ equal to the length of $r(n)$, we set $f(n) = 0$ when $n = 0$. Instantaneous frequency used in the paper is centered instantaneous frequency $F(n)$, which utilizes the mean of $f(n)$ to implement self-centralization and it can be represented as

$$F(n) = f(n) - \frac{1}{N}\sum_{n=1}^{N} f(n). \tag{7}$$

Instantaneous features are designed based on the instantaneous amplitude and the instantaneous frequency as follows.

- The number of instantaneous amplitude of received signals within a given range is defined as

$$K = \left\{ n : 0.4 < \left| \frac{r(n) - \mathrm{E}(r(n))}{\mathrm{std}(r(n))} \right| < 0.8 \right\}, \quad (8)$$

where $\mathrm{std}(\cdot)$ represents standard deviation operation. We normalize K to sampling points and we have

$$K_1 = \frac{K}{N}. \quad (9)$$

- The standard deviation of instantaneous amplitude is defined as

$$\sigma_a = \sqrt{\frac{1}{N}\left(\sum_{n=1}^{N} A^2(n)\right) - \left(\frac{1}{N}\sum_{n=1}^{N} A(n)\right)^2}. \quad (10)$$

- The standard deviation of absolute value of instantaneous amplitude is defined as

$$\sigma_{aa} = \sqrt{\frac{1}{N}\left(\sum_{n=1}^{N} A^2(n)\right) - \left(\frac{1}{N}\sum_{n=1}^{N} |A(n)|\right)^2}. \quad (11)$$

- The standard deviation of absolute value of instantaneous frequency is defined as

$$\sigma_{af} = \sqrt{\frac{1}{N}\left(\sum_{n=1}^{N} F^2(n)\right) - \left(\frac{1}{N}\sum_{n=1}^{N} |F(n)|\right)^2}. \quad (12)$$

- The kurtosis of instantaneous amplitude is defined as

$$\mu_a = \frac{\mathrm{E}\big(A^4(n)\big)}{\left[\mathrm{E}(A^2(n))\right]^2}. \quad (13)$$

- The kurtosis of instantaneous frequency is defined as

$$\mu_f = \frac{\mathrm{E}\big(F^4(n)\big)}{\left[\mathrm{E}(F^2(n))\right]^2}. \quad (14)$$

B. Definitions of Statistical Features

Statistical features include high-order moments and cumulants of the received signal.

- For a complex-valued signal $r(n)$, the k^{th}- order mixed moment with q conjugations $M_{k,q}$ is defined as

$$M_{k,q} = \mathrm{E}\big[r(n)^p \cdot \big(r(n)^*\big)^q\big], \quad (15)$$

where $p + q = k$, $r(n)^*$ is the conjugations of $r(n)$. Throughout the paper, the moments of interest are the 2^{nd}-order, 3^{th}-order, 4^{th}-order, 6^{th}-order, 8^{th}-order, 12^{th}-order and 16^{th}-order moments: $M_{2,0}$, $M_{2,1}$, $M_{3,0}$, $M_{3,1}$, $M_{4,0}$, $M_{4,1}$, $M_{4,2}$, $M_{6,0}$, $M_{6,1}$, $M_{6,2}$, $M_{6,3}$, $M_{8,0}$, $M_{8,1}$, $M_{8,2}$, $M_{8,3}$, $M_{8,4}$, $M_{12,0}$, $M_{12,1}$, $M_{12,2}$, $M_{12,3}$, $M_{12,4}$, $M_{12,5}$, $M_{12,6}$, $M_{16,0}$, $M_{16,1}$, $M_{16,2}$, $M_{16,3}$, $M_{16,4}$, $M_{16,5}$, $M_{16,6}$, $M_{16,7}$, and $M_{16,8}$.
- The cumulants are defined using joint cumulant function [50]. For example, a 6^{th}-order cumulant $C_{6,3}$ of $r(n)$ is defined as

$$C_{6,3} = \mathrm{cum}\big[r(n), r(n), r(n), r(n)^*, r(n)^*, r(n)^*\big], \quad (16)$$

where $\mathrm{cum}[\cdot]$ is the joint cumulant function, and the normalized 6^{th}-order cumulant $\widehat{C}_{6,q}$ is defined as

$$\widehat{C}_{6,q} = \frac{C_{6,q}}{\big(C_{4,2}\big)^3}, q = 0, 1, 2. \quad (17)$$

Besides, the normalized 4^{th}-order cumulant $\widehat{C}_{4,2}$ and the normalized 8^{th}-order cumulant $\widehat{C}_{8,q}$ are implemented with the following calculations:

$$\widehat{C}_{4,2} = \frac{C_{4,2}}{\big(C_{2,1}\big)^2}, \quad (18)$$

$$\widehat{C}_{8,q} = \frac{C_{8,q}}{\big(C_{4,2}\big)^2}, q = 0, 1, 2, 3. \quad (19)$$

In this paper, the 2^{nd}-order, 3^{rd}-order, 4^{th}-order, 6^{th}-order, 8^{th}-order cumulants as well as the normalized 4^{th}-order, 6^{th}-order and 8^{th}-order cumulants are used: $C_{4,0}$, $C_{4,1}$, $C_{4,2}$, $C_{6,0}$, $C_{6,1}$, $C_{6,2}$, $C_{6,3}$, $C_{8,0}$, $C_{8,1}$, $C_{8,2}$, $C_{8,3}$, $C_{8,4}$, $\widehat{C}_{4,2}$, $\widehat{C}_{6,0}$, $\widehat{C}_{6,1}$, $\widehat{C}_{6,2}$, $\widehat{C}_{8,0}$, $\widehat{C}_{8,1}$, $\widehat{C}_{8,2}$ and $\widehat{C}_{8,3}$.
- A new signal sequence $z(n, D)$ is defined which is calculated by

$$z(n, D) = r(n)r^*(n - D), \quad (20)$$

where D is an integer constant and $r^*(n - D)$ is the conjugate of $r(n - D)$. In this paper, we consider D to be the constant 2, 4 and 8. When $z(n, D)$ is obtained, we then calculate the above mentioned statistical features of $z(n, 2)$, $z(n, 4)$ and $z(n, 8)$, and splice them with the statistical features of $r(n)$.

C. Definitions of Spectral Features

Spectral features are designed according to the Fourier transform of the signal.

- The maximum value of spectral power density of instantaneous amplitude [31] is defined as

$$\gamma_{\max} = \max\Big(|\mathrm{DFT}(A(n))|^2\Big), \quad (21)$$

where $\mathrm{DFT}(\cdot)$ represents the discrete Fourier transform, i.e., $A(k) = \mathrm{DFT}(A(n)) = \sum_{n=0}^{N-1} A(n)e^{-j2\pi kn/N}$, $k = 0, 1, \ldots, N - 1$.
- The symmetry of spectrum [31] is to measure whether the spectrum is symmetric or asymmetric, which can be calculated through the following equation:

$$P = \frac{P_L - P_U}{P_L + P_U}, \quad (22)$$

where $P_L = \sum_{k=0}^{\frac{N}{2}-1} |R(k)|^2$, $P_U = \sum_{k=\frac{N}{2}}^{N-1} |R(k)|^2$, and $R(k) = \mathrm{DFT}(r(n))$.
- In order to take full use of spectral amplitude, we define an operation Find(x) as finding the three local maximum values of spectral amplitude. Thus, we have

$$\left[F_1^1, F_2^1, F_3^1\right] = \mathrm{Find}(10\lg|\mathrm{DFT}(r(n))|),$$

TABLE I
ADOPTED TRADITIONAL FEATURES

Categories	Variable	Features
Instantaneous Features	$r(n)$	K, σ_a, σ_{aa}, σ_{af}, μ_a, μ_f.
Statistical Features	$r(n)$	$M_{2,0}, M_{2,1}, M_{3,0}, M_{3,1}, M_{4,0}, M_{4,1}, M_{4,2}, M_{6,0}, M_{6,1}, M_{6,2}, M_{6,3}, M_{8,0}, M_{8,1}, M_{8,2}, M_{8,3}, M_{8,4}$
	$z(n,2)$	$M_{12,0}, M_{12,1}, M_{12,2}, M_{12,3}, M_{12,4}, M_{12,5}, M_{12,6}, M_{16,0}, M_{16,1}, M_{16,2}, M_{16,3}, M_{16,4},$
	$z(n,4)$	$M_{16,5}, M_{16,6}, M_{16,7}, M_{16,8}, C_{4,0}, C_{4,1}, C_{4,2}, C_{6,0}, C_{6,1}, C_{6,2}, C_{6,3}, C_{8,0}, C_{8,1}, C_{8,2},$
	$z(n,8)$	$C_{8,3}, C_{8,4}, \widehat{C}_{4,2}, \widehat{C}_{6,0}, \widehat{C}_{6,1}, \widehat{C}_{6,2}, \widehat{C}_{8,0}, \widehat{C}_{8,1}, \widehat{C}_{8,2}, \widehat{C}_{8,3}.$
Spectral Features	$r(n)$	$\gamma_{\max}, P, F_1^1, F_2^1, F_3^1, F_1^2, F_2^2, F_3^2, F_1^4, F_2^4, F_3^4, F_1^8, F_2^8, F_3^8.$

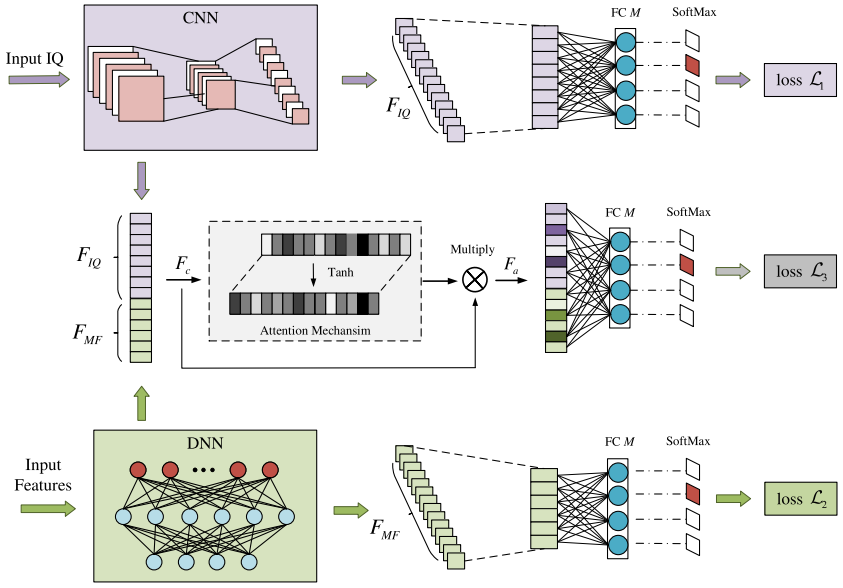


Fig. 1. The structure of the proposed HKDD framework. In this figure, "F_{IQ}" is the learned features by CNN and "F_{MF}" is the learned features from multiple traditional features. "FC M" is the final classification layer and M is equivalent to the number of categories to classify.

$$\begin{aligned}\left[F_1^2, F_2^2, F_3^2\right] &= \mathrm{Find}\left(10\lg\left|\mathrm{DFT}\left(r^2(n)\right)\right|\right),\\ \left[F_1^4, F_2^4, F_3^4\right] &= \mathrm{Find}\left(10\lg\left|\mathrm{DFT}\left(r^4(n)\right)\right|\right),\\ \left[F_1^8, F_2^8, F_3^8\right] &= \mathrm{Find}\left(10\lg\left|\mathrm{DFT}\left(r^8(n)\right)\right|\right). \quad (23)\end{aligned}$$

The logarithm operation is performed to avoid excessive values.

D. Summary of Adopted Features

In summary, the adopted traditional features in this paper are shown in Table I. It should be noted that we calculate 52 statistical features from each variable, $r(n)$, $z(n, 2)$, $z(n, 4)$ and $z(n, 8)$, respectively. From Table I, we can see that there are 6 instantaneous features, 208 statistical features and 14 spectral features. Thus, a total of 228 different features are adopted in this paper.

V. Proposed HKDD AMC Framework

In this section, we introduce the details of the proposed HKDD framework for modulation classification. First, a lightweight CNN consisting of depthwise separable convolution is constructed to deal with IQ sequence and a DNN with three hidden layers is used to deal with multiple traditional features. Then considering the varying importance of different features extracted from two different data sources, an attention mechanism is adopted in the HKDD network for learning the corresponding weight for each feature.

A. The Structure of HKDD Network

The HKDD network we designed contains two kinds of neural networks with different properties, CNN and DNN, because there are two kinds of data with different formats to be processed, i.e., the IQ data and the handcrafted features. Specifically, a CNN is designed for extracting features from IQ data, and a DNN with three fully connected layers is designed for processing the handcrafted features. For convenience, the traditional features are concatenated into a one-dimensional vector beforehand while the IQ data is shaped as $N \times 2$, where N represents the length of the IQ data. Details of the HKDD framework are shown in Fig. 1. Concatenation operation in

the HKDD is used to collect the learned features from both the CNN and DNN, which are represented by F_{IQ} and F_{MF} respectively in Fig. 1. F_{IQ} and F_{MF} are appended vertically to generate a joint feature representation F_c which is then sent to the module of attention mechanism to obtain the final feature vector F_a. Finally, F_a is sent to the last classification layer for obtaining probabilities of the received signal belonging to each modulation class. Besides concatenated to create a new feature vector, the features learned by CNN and DNN are also sent to the classification layers to predict results corresponding to their own inputs. Results of the three classification layers are used to update the parameters of the HKDD network in the training phase. However, in the inference phase, only the results obtained from F_a are used as the final classification results of our proposed HKDD.

B. CNN for IQ Input

Recently, some lightweight networks such as MobileNet [51], ShuffleNet [52] and Xception [53] have been proposed to diminish the network and speed up the training of the network through designing group convolution manually. The advantage of group convolution is that the parameters in the network can be greatly reduced. In this paper, when we consider the few-shot classification scenario, the imbalance between the large number of trainable parameters in the network and few labeled samples is a problem that needs to be addressed. Therefore, the use of group convolution can alleviate this problem to some extent.

The core depthwise separable convolution of MobileNet is a special group convolution, which can be divided into two steps, depthwise convolution and pointwise convolution. In the process of depthwise convolution, a single convolution kernel is used to convolve with a single channel of the input, which forces the number of kernels identical to the number of channels of the input. After the operation of depthwise convolution, the number of channels stays the same and the size of each channel may change due to downsampling. After the depthwise convolution, pointwise convolution is followed. The pointwise convolution is actually a conventional convolution with 1×1 kernels and the channels of output depend on the number of kernels explicated. The purpose of pointwise convolution is to ensure interchange of different feature maps because the corrections between channels are not taken into account during the process of depthwise convolution. Overall, the depthwise separable convolution, as shown in Fig. 2, can be expressed as

$$D_i = F_i \otimes K_i, i = 1, 2, \ldots, M, \tag{24}$$

$$F'_j = \sum_i D_i \otimes K_{j,i}, j = 1, 2, \ldots, P, \tag{25}$$

where D_i represents the i-th depthwise features after depthwise convolution, F_i is the i-th feature map of input, K_i is the convolution kernel of i-th group for depthwise convolution, $\otimes$ denotes the operation of convolution, F'_j represents the j-th output feature map after pointwise convolution and $K_{j,i}$ represents the j-th kernel with size 1×1 for pointwise convolution.

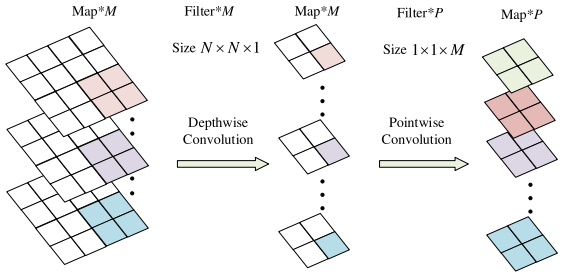


Fig. 2. Process of depthwise separable convolution. In this figure, "Map*M" means the number of feature maps is M and "Kernel*M" means the number of convolution kernels is M.

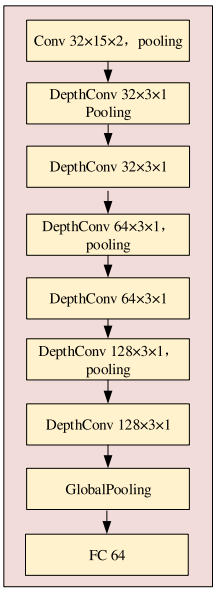


Fig. 3. The structure of lightweight CNN. In this figure, "conv $32 \times 15 \times 2$" denotes a convolutional layer with 32 kernels and the size of kernel is 15×2, "pooling" denotes that there is a maximum pooling layer after convolutional layer, and the stride of maximum pooling is 2, and "DepthConv" denotes the depthwise separable convolution. Batch normalization layer and activation layer between convolutional layer and maximum pooling layer are not shown for simplicity.

The depthwise separable convolution is used to design a lightweight CNN for extracting features from IQ data. The structure of the designed lightweight CNN is shown in Fig. 3. It mainly consists of one traditional convolution layer and six depthwise separable convolution layers. After obtaining feature maps from IQ data, a global pooling layer is used to transform the feature maps into a one-dimensional feature vector. Finally, a fully connected layer with 64 neurons is added to generate F_{IQ}.

C. DNN for Traditional Features Input

DNN is used to learn a mapping rule from input space $\mathcal{S}$ to target space $\mathcal{T}$ through a parametric function $\mathcal{F}_\theta : \mathcal{S} \to \mathcal{T}$, where parameters θ are specified by layers in DNN. The layers in DNN are called fully connected layers or dense layers. As a fundamental layer in DNN, the dense layer achieves the function of affine transformation:

$$f(x) = Wx + b, \tag{26}$$

where W and b are the trainable parameters in a single dense layer. For a DNN consisting of several dense layers, the parametric function $\mathcal{F}_\theta$ that maps the input to the desired output is presented by

$$\mathcal{F}_\theta = \circ f_n(\circ f_{n-1}(\cdots \circ f_1(x))), \tag{27}$$

where the parameters θ in $\mathcal{F}_\theta$ are a union of parameters in each dense layer. However, if a DNN is a simple composition of multiple dense layers, then $\mathcal{F}_\theta$ is unable to represent non-linear relation between the input and the output. For this reason, DNN applies activation layers to introduce non-linear function interleaved with dense layers. $\circ$ in (27) denotes the non-linear function.

In our proposed HKDD framework, the input space $\mathcal{S}$ of DNN consists of handcrafted features shaped as one-dimensional vectors and the target space $\mathcal{T}$ consists of the true labels corresponding to the input. As we need to focus on the hidden feature vector F_{MF}, the mapping rule from the input space to the output space is modified as $\mathcal{F}_\theta : \mathcal{S} \to F_{MF} \to \mathcal{T}$. Parametric function $f_{\theta'} : \mathcal{S} \to F_{MF}$ is specified by a DNN architecture composed of three dense layers interleaved with ReLU layers. The number of neurons in the three dense layers are 128, 96 and 64 respectively.

D. Attention Mechanism in HKDD Network

We represent F_{IQ} as $M_i = [m_1, m_2, \ldots, m_i]$ and F_{MF} as $N_j = [n_1, n_2, \ldots, n_j]$. The joint feature representation F_c is obtained by combining the F_{IQ} and the F_{MF}, which can be represented as follows:

$$F_c = M_i \oplus N_j, \tag{28}$$

where the operation of $\oplus$ is implemented as a concatenate function. Thus, $F_c = [m_1, m_2, \ldots, m_i, n_1, n_2, \ldots, n_j]$, $F_c \in \mathbb{R}^{i+j}$. It should be pointed out that not all features in F_c are helpful for classification. Some immature and adverse features may exist due to the inadequate learning in the few-shot scenario. Thus, we adopt an attention mechanism to learn corresponding weight vector W_c, $W_c \in \mathbb{R}^{i+j}$ which is obtained from a DNN with three fully connected layers. The number of neurons in the first and last fully connected layer is identical to the number of features in F_c while the number of neurons in the second fully connected layer is half of the number of features in F_c. The activation functions of these fully connected layers are Tanh and Sigmoid. After the learned weights are finally activated by Sigmoid, their values are forced to distribute between 0 and 1. In this way, the immature features could be abandoned through assigning their weights with small values. On the contrary, the important features' weights will be assigned with values closed to 1.

The process of the attention mechanism can be represented as follows:

$$\begin{aligned} Q_1 &= \mathrm{Tanh}(W_1 F_c + b_1), \\ Q_2 &= \mathrm{Tanh}(W_2 Q_1 + b_2), \\ W_c &= \mathrm{Sigmoid}(W_3 Q_2 + b_3), \\ F_a &= F_c \otimes W_c, \end{aligned} \tag{29}$$

where $W_1 \in \mathbb{R}^{(i+j)\times(i+j)}$, $W_2 \in \mathbb{R}^{(i+j)\times(i+j)/2}$, $W_3 \in \mathbb{R}^{(i+j)/2\times(i+j)}$ are the trainable weight matrices, $b_1, b_3 \in \mathbb{R}^{(i+j)}$, $b_2 \in \mathbb{R}^{(i+j)/2}$ are the trainable biases. $\mathrm{Tanh}(\cdot)$ and $\mathrm{Sigmoid}(\cdot)$ denote the tanh activation function and sigmoid activation function respectively. Multiplication $\otimes$ is defined as $F_{a,t} = F_{c,t} \cdot W_{c,t}, t = 1, 2, \ldots, (i+j)$. The joint feature representation F_c is transferred to F_a after using the attention mechanism.

E. Loss Function for HKDD Network

The goal of training is to optimize the weights and biases of the network by minimizing the loss between the true output of the network and the desired output or the given label of training data. In the supervised training process, the output of the network is a probability distribution with respect to categories of the classification problem. As a commonly used loss function, cross-entropy is adopted to measure the error between true probability distribution $P(x) = [p(x_1), p(x_2), \ldots, p(x_M)]$ and predicted probability distribution $Q(x) = [q(x_1), q(x_2), \ldots, q(x_M)]$, which can be represented as

$$\mathcal{L} = -\sum_{i=1}^{M} p(x_i) \log(q(x_i)), \tag{30}$$

where M is the number of categories designed to classify, $p(x_i)$ represents the true probability belonging to the i-th class and $q(x_i)$ represents the predicted probability belonging to the i-th class. However, in the HKDD network, there are three predicted probability distribution $Q_1(x) = [q_1(x_1), q_1(x_2), \ldots, q_1(x_M)]$, $Q_2(x) = [q_2(x_1), q_2(x_2), \ldots, q_2(x_M)]$ and $Q_3(x) = [q_3(x_1), q_3(x_2), \ldots, q_3(x_M)]$, which are obtained from the predicted results of F_{IQ}, F_{MF} and the joint feature representation F_c. So the loss function in the HKDD network can be represented as

$$\mathcal{L}_{sum} = -\sum_{m=1}^{3}\sum_{i=1}^{M} p(x_i) \log(q_m(x_i)). \tag{31}$$

$\mathcal{L}_{sum}$ is used to update parameters in the HKDD network and Adam optimizer is adopted during training.

The weights and biases are adjusted iteratively by applying the gradient of loss, which is given as:

$$w_k^{l+1}(i+1) = w_k^{l+1}(i) - \eta \nabla \mathcal{L} = w_k^{l+1}(i) - \eta \frac{\partial \mathcal{L}}{\partial w_k^{l+1}(i)}, \tag{32}$$

$$b_k^{l+1}(i+1) = b_k^{l+1}(i) - \eta \nabla \mathcal{L} = b_k^{l+1}(i) - \eta \frac{\partial \mathcal{L}}{\partial b_k^{l+1}(i)}, \tag{33}$$

where w_k^{l+1} represents the trainable k-th weight in $(l+1)$-th layer while b_k^{l+1} is the corresponding bias, and η is the step size.

In summary, the training algorithm for HKDD network is shown in Algorithm 1.

VI. Simulation Results

In this section, we first give the settings for simulation, which include the two datasets we build and parameter settings

Algorithm 1 The Training Process of the HKDD Network

Input: IQ training set $\mathcal{D} = \left\{\left([I(n), Q(n)]^{(i)}, y^{(i)}\right)\right\}_{i=1}^{\mathcal{N}}$, features training set $\mathcal{F} = \left\{F^{(i)}, y^{(i)}\right\}_{i=1}^{\mathcal{N}}$, minibatch N_b, learning rate r, maximum number of iterations $\mathcal{I}$;

Randomly initialize parameters of the network;

for $t = 1, 2, \ldots, \mathcal{I}$ **do**

1. Randomly choose a mini-batch of N_b samples from $\mathcal{D}$ and $\mathcal{F}$ with the same sample indexes;

2. Calculate the total classification loss according to (31);

3. Update parameters of the network according to (32) and (33);

end for

Output: The HKDD model f_θ.

for training. Then we provide the simulation results in two scenarios: the adequate-sample scenario with enough training samples and the few-shot scenario with a small number of training samples. Finally, we compare the performance of our proposed method with existing AMC methods.

A. Simulation Settings

1) Datasets: In order to verify the performance of our proposed modulation classification method, we generate two different datasets using MATLAB.[1]

- *HKDD_AMC12:* A dataset containing 12 different modulation signals: BPSK, QPSK, 8PSK, OQPSK, 16QAM, 32QAM, 64QAM, 4PAM, 8PAM, 2FSK, 4FSK, 8FSK.
- *HKDD_AMC36:* A dataset containing 36 different modulation signals: BPSK, QPSK, 8PSK, OQPSK, 16PSK, 32PSK, 2FSK, 4FSK, 8FSK, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM, 16APSK, 32APSK, 64APSK, 128APSK, 256APSK, 4PAM, 8PAM, 16PAM, MSK, GMSK, 4CPM, 8CPM, OFDM-BPSK, OFDM-QPSK, OFDM-16QAM, AM, FM, OOK, 4ASK, 8ASK, AM-MSK, FM-MSK.

In both datasets, the original bit sequence is chosen from 0 and 1 in a random manner to ensure that the probability of appearance for each symbol is equivalent. The length of each modulated signal is 1024 for dataset HKDD_AMC36 and 512 for dataset HKDD_AMC12. The oversampling rate is 8, so each sampled sequence in dataset HKDD_AMC36 contains 128 symbols and each sampled sequence in dataset HKDD_AMC12 contains 64 symbols. A root raised-cosine (RRC) filter with 6-symbols truncated length is employed as the pulse-shaping filter and the roll-off coefficient of RRC is randomly chosen within the range 0.2 to 0.7. The frequency offset is randomly chosen from -0.2 to 0.2 (normalized to the sampling frequency). The range of SNR is (-20 dB, 30 dB) for dataset HKDD_AMC36 and (-20 dB, 20 dB) for dataset HKDD_AMC12 with an interval of 2 dB. The number of training samples for each modulation type is 1000 in each SNR and the number of testing samples is half of the training samples. Both datasets contain both IQ signals and traditional features. HKDD_AMC36 is the dataset currently available that contains the most number of modulation categories.

2) Model Training: The datasets mentioned above are separately used to train the HKDD network. For dataset HKDD_AMC36, the HKDD network is regarded as a classifier with 36 categories, while for dataset HKDD_AMC12, the output category of the HKDD network is 12. In the process of training, the mini-batch size is 128 for HKDD_AMC12 and 256 for HKDD_AMC36. The total number of parameters of the proposed HKDD network is about 0.1 M, about 0.03 M for the DNN part and about 0.07 M for the CNN part. In training HKDD, the initial learning rate is 0.003 and after every 5 epochs, the learning rate is reduced to half of the previous value. The network is trained for 30 epochs, which takes about half an hour with NVIDIA GeForce RTX 2080.

B. Performance in Adequate-Sample Scenario

We first verify the performance of the proposed HKDD on the two datasets. For comparison, the performance of DNN using traditional features (denoted as the DNNTF method) and the CNN using IQ sequence (denoted as the CNNIQ method) on the two datasets is also given. Fig. 4(a) shows the performance of the methods on the dataset HKDD_AMC12. We can see that in the low SNR region, the CNNIQ method gets the worst performance. The performance of the DNNTF method is better than the CNNIQ method because the traditional features used by the DNNTF method, such as σ_{af}, high-order moments and cumulants, are helpful to distinguish M-PSK and M-FSK signals in low SNR region. However, when the SNR increases, the performance of the DNNTF method is inferior to the CNNIQ method. This is because CNN has the ability to extract deeper-level features compared with the traditional features used by the DNNTF method. Because of the two methods' respective different contributions, the HKDD network achieves the best performance in all SNR ranges. Fig. 4(b) illustrates the performance of these methods on the dataset HKDD_AMC36 which contains 36 modulation types. We can see that the HKDD network still achieves the best performance, with about 8% absolute accuracy improvement compared with CNNIQ in the low SNR region and about 10% absolute accuracy improvement compared with the DNNTF method in the high SNR region.

In order to evaluate the performance of the attention mechanism, the performance of the HKDD network without the attention mechanism is also given. It can be observed that in this adequate-sample scenario, the HKDD method without the attention mechanism achieves nearly the same performance as that of the HKDD method. This is because the network is fully trained and the attention mechanism does not provide additional information in this adequate-sample scenario. To further illustrate the function of the attention mechanism, we plot the weight vectors W_c of attention mechanism in Fig. 5. The values in W_c are used to measure the extent of importance for the features in F_a. We test on 128 samples and the obtained weight vectors, each of which is with length of

[1]The datasets are available at https://github.com/yexijoe/HKDD.

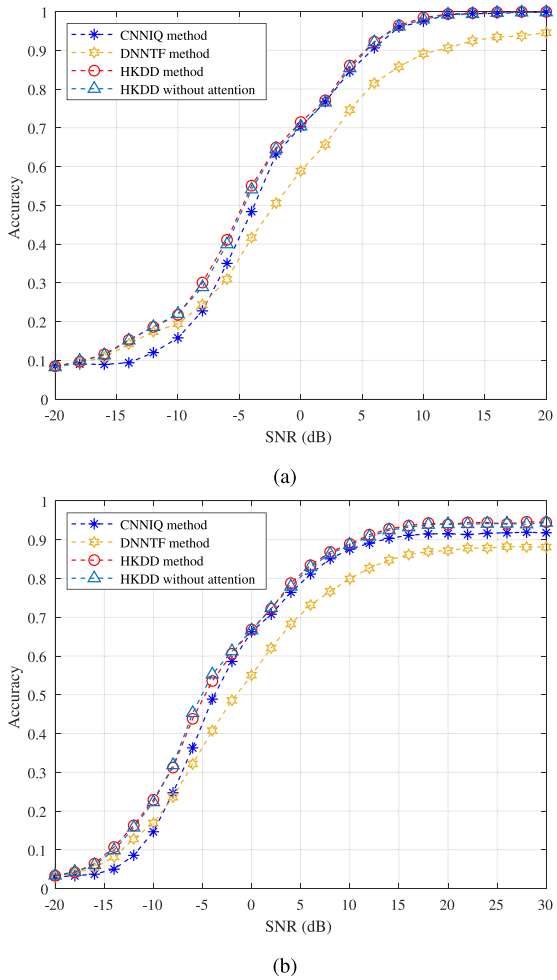


Fig. 4. Performance of HKDD network on the two datasets. (a) HKDD_AMC12 and (b) HKDD_AMC36.

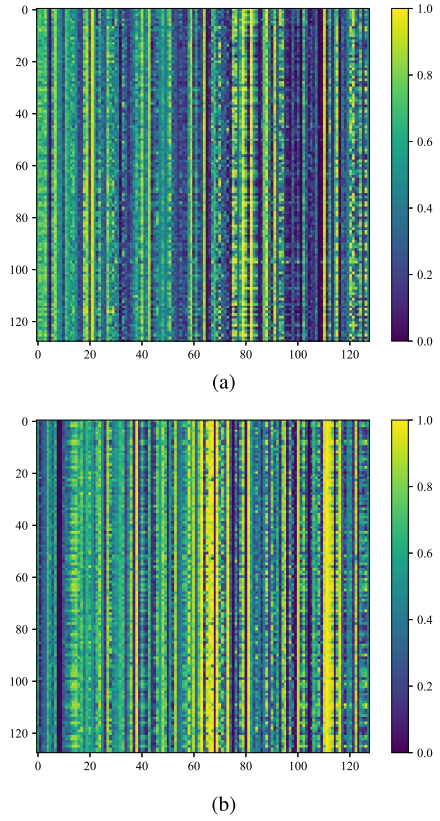


Fig. 5. Weights of attention mechanism for joint features. (a) Dataset HKDD_AMC12 and (b) dataset HKDD_AMC36.

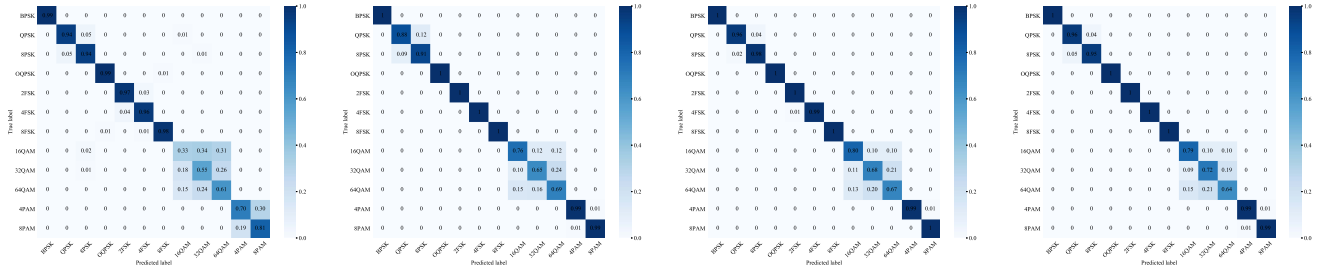
128 for each sample, form a 128 × 128 matrix with values distributed between 0 and 1. In Fig. 5(a) and Fig. 5(b), the left half shows the weights corresponding to the features learned by CNN while the right half illustrates the weights corresponding to the features learned by DNN. We can see that the two parts of the weight vectors are relatively average, which shows that the features learned from raw IQ and the features learned from traditional features are both important for AMC in adequate-sample scenario.

For an intuitive presentation, we further provide the confusion matrices of classification on HKDD_AMC12 and HKDD_AMC36, which are shown in Fig. 6 and Fig. 7 respectively. In Fig. 6(a), the confused modulations for DNNTF method are among 16QAM, 32QAM and 64QAM, and between 4PAM and 8PAM. It is clear that 16QAM exhibits the worst classification accuracy of about 33%, which is confused with 32QAM by 34% and with 64QAM by 31%. In Fig. 6(b), with CNNIQ method, 16QAM, 32QAM and 64QAM are also confused. Furthermore, QPSK is confused with 8PSK. However, in Fig. 6(c), for our proposed HKDD method, the confusion between 4PAM and 8PAM, as well as the confusion between QPSK and 8PSK do not exist. What's more, the classification of 16QAM, 32QAM and 64QAM is improved in Fig. 6(c), which illustrates that our proposed method is effective by incorporating the knowledge into the DL model. Fig. 6(d) shows the confusion matrix of HKDD without attention mechanism, which is similar to that of HKDD in Fig. 6(c) as expected.

In Fig. 7, for clarity of graphical representation, we split the confusion matrix of the 36-modulation classification into a confusion matrix of 19-modulation classification and a confusion matrix of 17-modulation classification. For example, Fig. 7(a) and Fig. 7(e) together represent the classification confusion of DNNTF method on HKDD_AMC36. We can see from Fig. 7 that the main classification confusions arise in high-order PSK modulations, such as 8PSK, 16PSK, and 32PSK, high-order QAM modulations, such as 64QAM, 128QAM, 256QAM, and OFDM modulations, i.e., OFDM-QPSK and OFDM-16QAM. Similarly, in Fig. 7(c) and Fig. 7(g), these classification confusions are ameliorated when using the HKDD method.

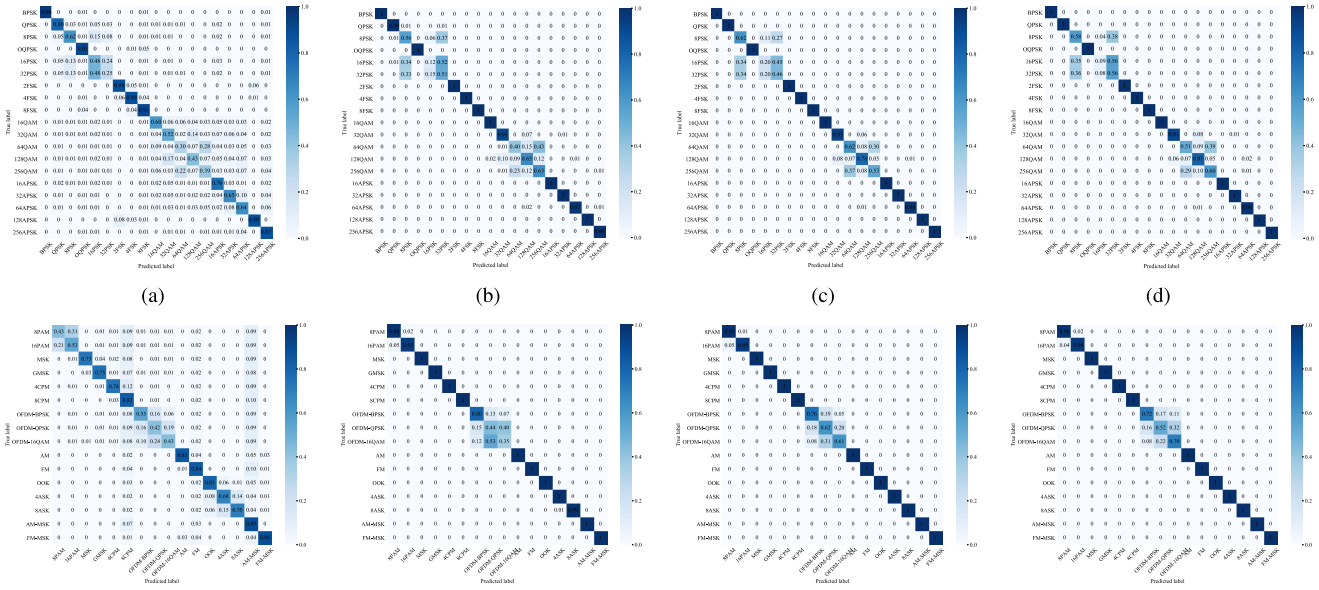
C. Performance in Few-Shot Scenario

We now discuss the influence of few-shot learning to the performance of the HKDD network. For dataset HKDD_AMC12, four few-shot scenarios are considered: (1) 10% of samples, (2) 5% of samples, (3) 1% of samples and (4) 0.5% of samples. For dataset HKDD_AMC36, the four few-shot scenarios considered are: (1) 5% of samples, (2) 1% of samples, (3) 0.5% of samples and (4) 0.2% of samples.



(a) (b) (c) (d)

Fig. 6. Confusion matrices of modulation classification at 6 dB for dataset HKDD_AMC12 with different methods. (a) DNNTF method, (b) CNNIQ method, (c) proposed HKDD and (d) HKDD without attention.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 7. Confusion matrices of modulation classification at 10 dB for dataset HKDD_AMC36 with different methods. (a) DNNTF method (19-modulation), (b) CNNIQ method (19-modulation), (c) HKDD method (19-modulation), (d) HKDD without attention (19-modulation), (e) DNNTF method (17-modulation), (f) CNNIQ method (17-modulation) (g) HKDD method (17-modulation) and (h) HKDD without attention (17-modulation).

Considering that the network will not have enough opportunity to update the parameters as the number of training samples decreases, we reduce the batch size to alleviate this situation. Batch size in each few-shot scenario is set to 64, 64, 36, 36 for dataset HKDD_AMC12 and 128, 96, 64, 48 for dataset HKDD_AMC36.

Fig. 8 shows the performance of the three methods in four few-shot scenarios for dataset HKDD_AMC12. The performance of the HKDD network without the attention mechanism is also given. Compared with the simulation results in Fig. 8(a), we can see that when 10% samples are used, the decline in classification accuracy of the CNNIQ method is more dramatic than that of the DNNTF method. Specifically, the classification accuracy of the CNNIQ method drops from nearly 100% to about 78% while the classification accuracy of the DNNTF method drops from around 95% to around 92%. In Fig. 8(d), we can see that the most obvious trend is that with the decrease in the number of training samples, the classification accuracy of the CNNIQ method rapidly declines. On the contrary, the classification accuracy of the DNNTF method is only slightly reduced. To be more specific, the accuracy of the CNNIQ method is decreased from about 78% to around 43% and the accuracy of the DNNTF method is decreased from about 92% to around 83% in the highest SNR, i.e., 20 dB in this case. It means that the CNNIQ method which needs a large number of training samples to learn a mapping from input to output has inferior performance on the task of few-shot classification compared with the DNNTF method. That is because, during the training of CNN, overfitting will occur in the few-shot scenario, and the fewer the training samples, the more serious the overfitting. However, the traditional features used in the DNNTF method represent the low-dimensional information of signals. Unlike the high-dimensional information extracted by CNN, which requires a large number of samples to learn, the low-dimensional features can directly obtain the result of classification through several fully connected layers. Nevertheless, the HKDD network can achieve remarkable performance gain when 10% and 5% of samples are used. Although the CNNIQ method performs poorly in the scenario of 0.5% of samples, the HKDD network

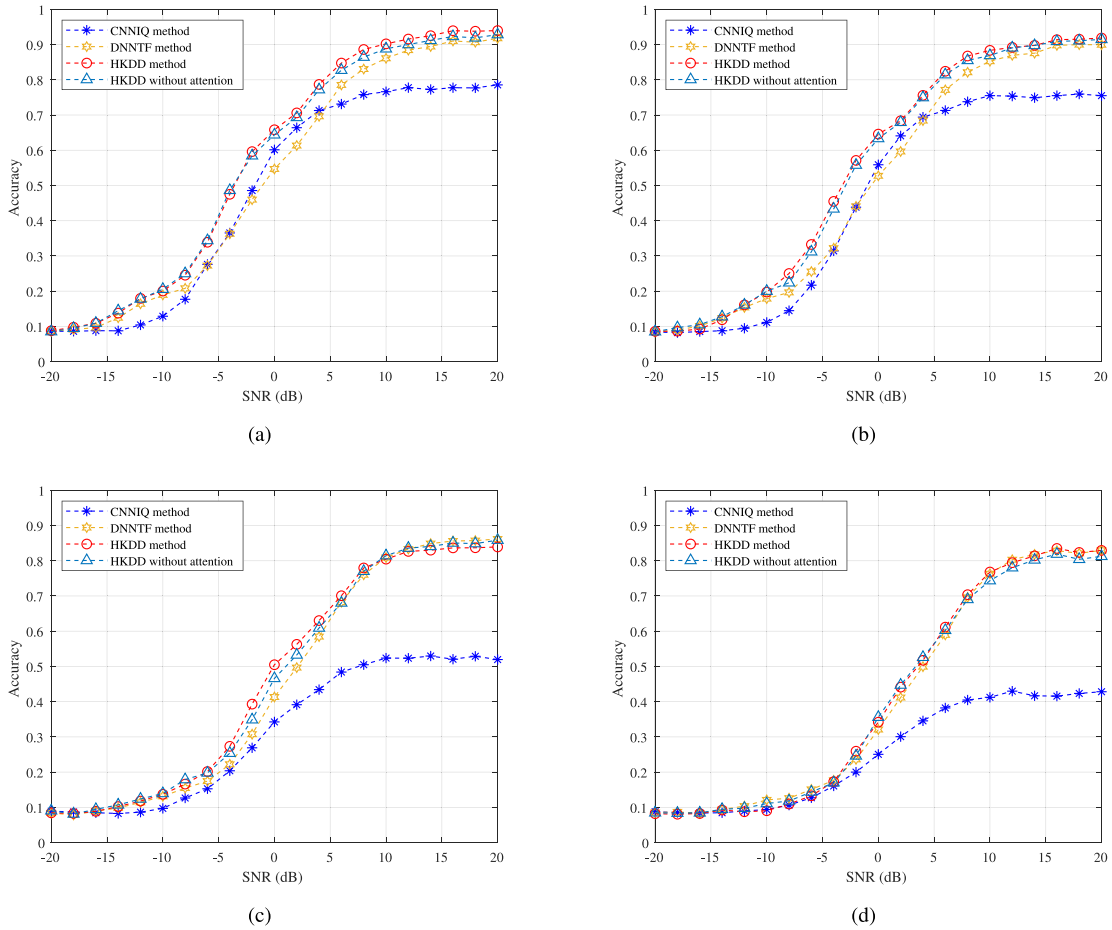


Fig. 8. Performance of HKDD network in few-shot scenarios for dataset HKDD_AMC12. (a) 10% of samples, (b) 5% of samples, (c) 1% of samples and (d) 0.5% of samples.

can still achieve the same performance as the DNNTF method. It is the result of the attention mechanism, which will discard a part of features extracted by CNN and pay more attention to the features of DNN as we will discuss later. Besides, we notice that when the attention mechanism is not added, the HKDD suffers from a little performance loss.

We further verify the effectiveness of the HKDD network for the few-shot classification on dataset HKDD_AMC36. The classification performance is shown in Fig. 9. It is similar to the trend in Fig. 8, that is, as the number of training samples decreases, the classification accuracy of the CNNIQ method declines faster than that of the DNNTF method. In the four few-shot scenarios, the difference between the highest accuracy and the lowest accuracy is about 41% for the CNNIQ method, and about 15% for the DNNTF method. It should be noted that the SNR range is -20 dB to 30 dB in this experiment, which is wider than that of the above experiment. For this reason, the number of samples used in this experiment is actually more when the same proportion of dataset samples are used for the two experiments. So, we consider an extreme situation where only 2 samples are used in each SNR, and the simulation result is shown in Fig. 9(d). We can see that the DNNTF method achieves about 70% accuracy for classifying 36 modulation signals in the highest SNR, 30 dB in this case. The accuracy of the HKDD network is still a little higher than that of the DNNTF method. From Fig. 9, it can be found that the HKDD network is always the most effective method compared with the CNNIQ and DNNTF methods in terms of classification accuracy. However, when we remove the attention mechanism in the HKDD network, the classification accuracy of the HKDD network will decrease, which illustrates the attention mechanism is helpful in this framework.

To further explain the function of the attention mechanism, we draw the weights of the attention mechanism in Fig. 10. Similarly, we test on 128 samples and the corresponding weights for each sample form a vector with 128 values distributed between 0 and 1. The values are used to measure the extent of importance for the features in F_a. Fig. 10(a) shows the feature vector in attention mechanism when 1% of samples in dataset HKDD_AMC12 are used and Fig. 10(b) shows the feature vector when 1% of samples in dataset HKDD_AMC36 are used. In Fig. 10(a) and Fig. 10(b), the left half shows the weights corresponding to the features learned by CNN while the right half shows the weights corresponding to the features learned by DNN. We can see that few features extracted by CNN are allocated with weights close to 1. On the contrary, most weights for features extracted by DNN are with value close to 1. It illustrates that CNNIQ is vulnerable to

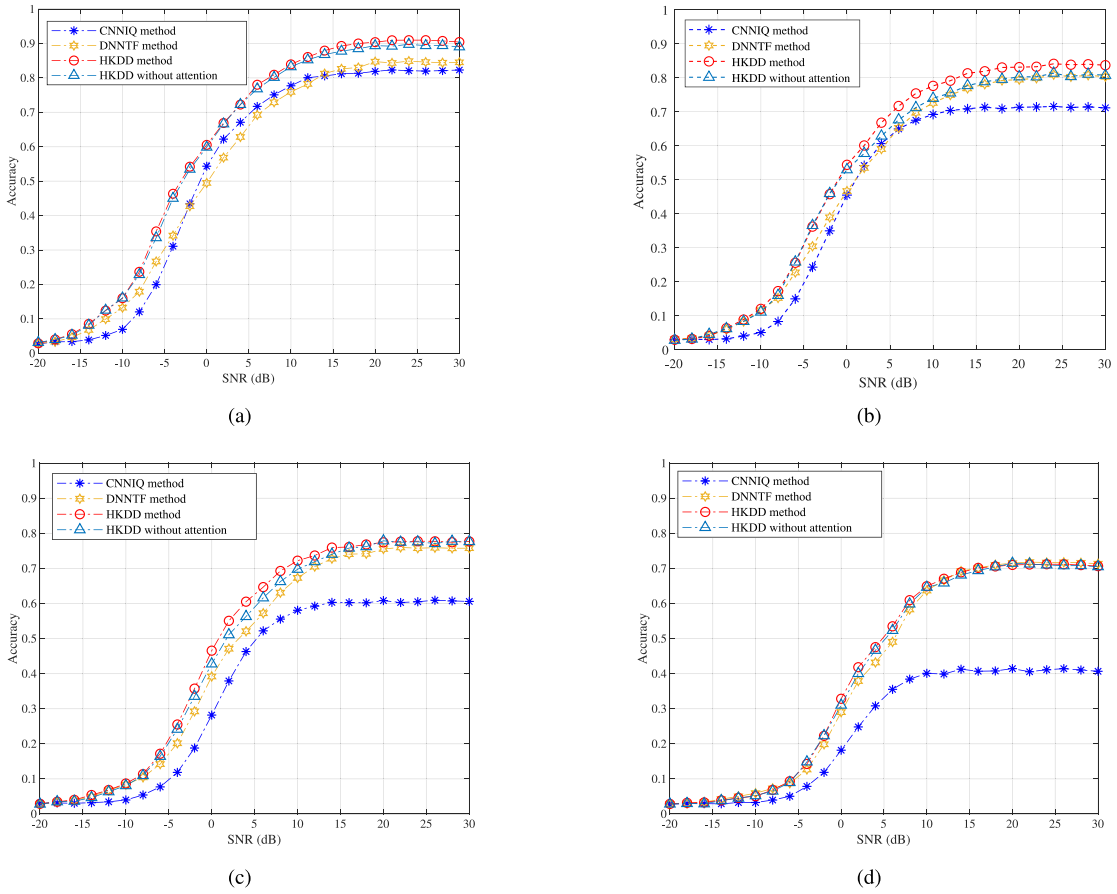


Fig. 9. Performance of HKDD network in few-shot scenarios for dataset HKDD_AMC36. (a) 5% of samples, (b) 1% of samples, (c) 0.5% of samples and (d) 0.2% of samples.

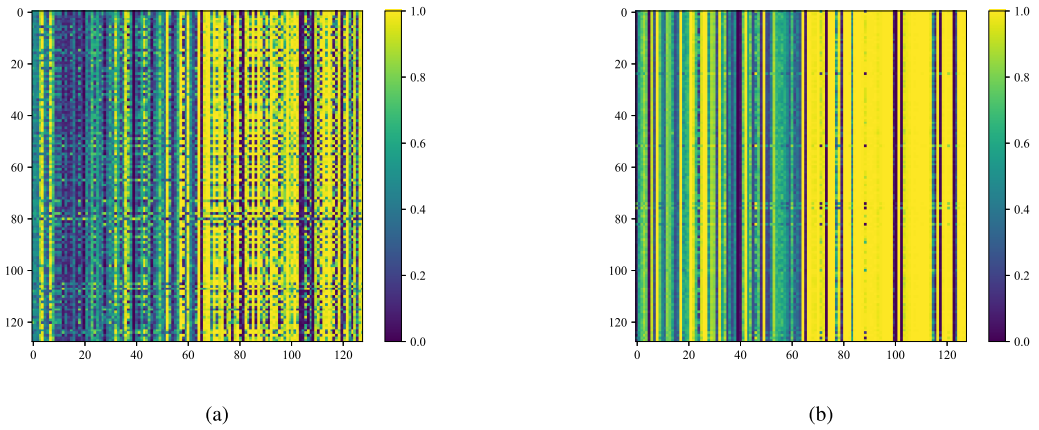
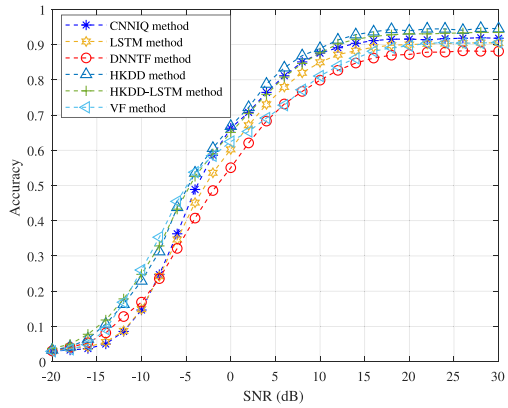


Fig. 10. Weights of attention mechanism for joint features. (a) 1% of samples in dataset HKDD_AMC12 and (b) 1% of samples in dataset HKDD_AMC36.

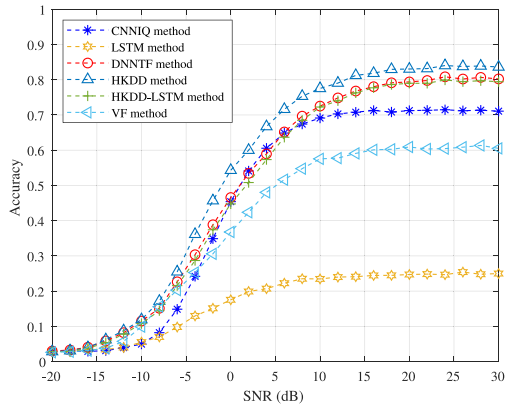
inadequate learning in few-shot scenario, and most features extracted by CNN are immature and detrimental to the classification results. The attention mechanism tends to highlight the features extracted by DNN due to the better performance of the DNNTF method in few-shot scenarios.

D. Comparison With Other AMC Methods

Finally, we compare the performance of the proposed HKDD with other AMC methods on dataset HKDD_AMC36. The existing AMC methods used for comparison include a raw IQ-based method which uses LSTM as the deep neural network structure and the hybrid VF method given in [48]. Specifically, the LSTM model used is a structure with 2 LSTM layers. The VF method uses time-frequency distribution instead of the raw IQ as one of its inputs which is quite different from our proposed HKDD. Fig. 11(a) and Fig. 11(b) show the comparison of modulation classification results on the complete dataset HKDD_AMC36 and it's few-shot scenario with only 1% training samples, respectively.



(a)



(b)

Fig. 11. Comparison with other AMC methods. (a) Complete dataset HKDD_AMC36 and (b) 1% of training samples of dataset HKDD_AMC36.

From Fig. 11(a), we can see that with sufficient training samples, the classification accuracy of LSTM in low SNR region is about the same as that of CNNIQ, both lower than that of DNNTF. In the high SNR region, the performance of LSTM is better than that of DNNTF though worse than that of CNNIQ. We also replace CNNIQ in HKDD with LSTM to get another hybrid version HKDD-LSTM. Obviously, compared with LSTM and DNNTF, the classification accuracy of HKDD-LSTM in all SNR range is greatly improved, which further shows the effectiveness of our proposed hybrid framework. As for the existing hybrid framework VF [48], in the low SNR region, the performance of VF is slightly better than that of HKDD, possibly by virtue of its usage of time-frequency distribution which is beneficial in low SNR. However, in high SNR region, the VF method suffers seriously and it performs even worse than CNNIQ. This may be caused by the lack of using raw IQ input as one of its inputs which is probability a severe limitation of VF. The performance gap between our proposed HKDD over the VF method is quite large in the high SNR region. In general, our proposed hybrid framework HKDD performs far better than VF.

Fig. 11(b) illustrates the results in the case of few-shot scenario. It is obvious that LSTM has the worst performance when there are few training samples. Meanwhile, HKDD-LSTM which combines LSTM and DNNTF improve the performance to close to that of the DNNTF method. In particular, in the case of few-shot scenario, the classification performance of VF in all SNR is inferior to that of HKDD. The performance gap is larger in the high SNR region. Specifically, when SNR $=$ 30 dB, the classification accuracy of VF is about 61%, however, the classification accuracy of HKDD is about 84%. It confirms that our proposed HKDD has excellent modulation recognition capability in few-shot scenario.

VII. Conclusion

In this paper, we have presented a HKDD framework for AMC which combines the knowledge-based method and the DL-based method in order to improve the classification performance in both the adequate-sample scenario the few-shot scenario. To take full advantage of the knowledge-based method, we have calculated various instantaneous features, statistical features and spectral features from the raw signal. In our HKDD framework, a CNN is used to extract features from IQ sequences and a DNN is used to process the handcrafted features. Moreover, a fusion method is adopted to combine learned features to form a joint feature vector and an attention mechanism is designed to abandon immature features and highlight important features. For validating the effectiveness of our proposed method, we have constructed two modulation classification datasets containing both traditional features and raw IQ data. Simulation results have proved that our proposed HKDD is superior to the DL-based method and the knowledge-based method in both scenarios. The performance gain increases remarkably with the decrease of the number of training samples. In addition, the attention mechanism has been proved to be useful in detecting the importance of different features when training in the few-shot scenario.

While recently data-driven deep learning plays an increasingly important role in AMC, we also pay our attention to traditional features derived from domain knowledge. The proposed HKDD framework focuses on the integration of the knowledge domain and the data domain which brings inspiration to build the next-generation signal intelligence. In the future work, we will extend HKDD to deal with other tasks in radio signal processing including signal sensing, signal parameter estimation, and specific emitter recognition.

References

[1] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 16–32, Jan. 2020.

[2] M. Bande, A. Magesh, and V. V. Veeravalli, "Dynamic spectrum access using stochastic multi-user bandits," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 953–956, May 2021.

[3] V. Aswathi and A. V. Babu, "Performance analysis of NOMA-based underlay cognitive radio networks with partial relay selection," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4615–4630, May 2021.

[4] H. Wang, J. Wang, G. Ding, J. Chen, Y. Li, and Z. Han, "Spectrum sharing planning for full-duplex UAV relaying systems with underlaid D2D communications," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1986–1999, Sep. 2018.

[5] A. O. A. Salam, R. E. Sheriff, Y.-F. Hu, S. R. AI-Araji, and K. Mezher, "Automatic modulation classification using interacting multiple model Kalman filter for channel estimation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8928–8939, Sep. 2019.

[6] S. Huang, Y. Jiang, X. Qin, Y. Gao, Z. Feng, and P. Zhang, "Automatic modulation classification of overlapped sources using multi-gene genetic programming with structural risk minimization principle," *IEEE Access*, vol. 6, pp. 48827–48839, 2018.

[7] O. A. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "Survey of automatic modulation classification techniques: Classical approaches and new trends," *IET Commun.*, vol. 1, pp. 137–156, Apr. 2007.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[9] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*.

[10] S. Zheng, S. Chen, P. Qi, H. Zhou, and X. Yang, "Spectrum sensing based on deep learning classification for cognitive radios," *China Commun.*, vol. 17, no. 2, pp. 138–148, Feb. 2020.

[11] S. Chen, S. Zheng, L. Yang, and X. Yang, "Deep learning for large-scale real-world ACARS and ADS-B radio signal classification," *IEEE Access*, vol. 7, pp. 89256–89264, 2019.

[12] Y. Tu et al., "Large-scale real-world radio signal recognition with deep learning," *Chin. J. Aeronaut.*, vol. 35, no. 9, pp. 35–48, 2022.

[13] S. Zheng, S. Chen, and X. Yang, "DeepReceiver: A deep learning-based intelligent receiver for wireless communications in the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 5–20, Mar. 2021.

[14] X.-H. Li et al., "A survey of data-driven and knowledge-aware eXplainable AI," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 29–49, Jan. 2022.

[15] J. Feng, Y. Yao, S. Lu, and Y. Liu, "Domain knowledge-based deep-broad learning framework for fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 68, no. 4, pp. 3454–3464, Apr. 2021.

[16] W. Yang, L. Jin, Z. Xie, and Z. Feng, "Improved deep convolutional neural network for online handwritten Chinese character recognition using domain-specific knowledge," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, 2015, pp. 551–555.

[17] X. Ding, Y. Chen, Z. Tang, and Y. Huang, "Camera identification based on domain knowledge-driven deep multi-task learning," *IEEE Access*, vol. 7, pp. 25878–25890, 2019.

[18] M. M. Shakra, E. M. Shaheen, H. A. Bakr, and M. S. Abdel-Latif, "C3. Automatic digital modulation recognition of satellite communication signals," in *Proc. 32nd Nat. Radio Sci. Conf. (NRSC)*, 2015, pp. 118–126.

[19] A. K. Nandi and E. E. Azzouz, "Algorithms for automatic modulation recognition of communication signals," *IEEE Trans. Commun.*, vol. 46, no. 4, pp. 431–436, Apr. 1998.

[20] M. A. Hazar, N. Odabasioglu, T. Ensari, Y. Kavurucu, and O. F. Sayan, "Performance analysis and improvement of machine learning algorithms for automatic modulation recognition over Rayleigh fading channels," *Neural Comput. Appl.*, vol. 29, no. 9, pp. 351–360, 2018.

[21] S.-Z. Hsue and S. S. Soliman, "Automatic modulation recognition of digitally modulated signals," in *Proc. IEEE Mil. Commun. Conf. Bridging Gap. Interoperability Survivability Security*, vol. 3, 1989, pp. 645–649.

[22] S.-Z. Hsue and S. S. Soliman, "Automatic modulation classification using zero crossing," *IEE Proc. F, Radar Signal Process.*, vol. 137, no. 6, pp. 459–464, 1990.

[23] L.-L. Meng and X.-J. Si, "An improved algorithm of modulation classification for digital communication signals based on wavelet transform," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, 2007, pp. 1226–1231.

[24] L. Zhang, Z. Yang, and W. Lu, "Digital modulation classification based on higher-order moments and characteristic function," in *Proc. IEEE 5th Int. Conf. Signal Image Process. (ICSIP)*, 2020, pp. 809–812.

[25] Y. Zhao, Y.-T. Xu, H. Jiang, Y.-J. Luo, and Z.-W. Wang, "Recognition of digital modulation signals based on high-order cumulants," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2015, pp. 1–5.

[26] H.-C. Wu, M. Saquib, and Z. Yun, "Novel automatic modulation classification using cumulant features for communications via multipath channels," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3098–3105, Aug. 2008.

[27] M. Sreekantamurthy, D. C. Popescu, and R. P. Joshi, "Classification of digital modulation schemes in multipath environments using higher order statistics," in *Proc. SoutheastCon*, 2016, pp. 1–6.

[28] D. Das, A. Anand, P. K. Bora, and R. Bhattacharjee, "Cumulant based automatic modulation classification of QPSK, OQPSK, π/4-QPSK and 8-PSK in MIMO environment," in *Proc. Int. Conf. Signal Process. Commun. (SPCOM)*, 2016, pp. 1–5.

[29] R. Gupta, S. Majhi, and O. A. Dobre, "Design and implementation of a tree-Based blind modulation classification algorithm for multiple-antenna systems," *IEEE Trans. Instrum. Meas.*, vol. 68, pp. 3020–3031, 2019.

[30] M. Abdelbar, W. H. Tranter, and T. Bose, "Cooperative cumulants-based modulation classification in distributed networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 446–461, Sep. 2018.

[31] A. K. Nandi and E. E. Azzouz, "Automatic analogue modulation recognition," *Signal Process.*, vol. 46, no. 2, pp. 221–222, 1995.

[32] P. C. Sapiano, J. D. Martin, and R. J. Holbeche, "Classification of PSK signals using the DFT of phase histogram," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 3, 1995, pp. 1868–1871.

[33] C. Schreyogg and J. Reichert, "Modulation classification of QAM schemes using the DFT of phase histogram combined with modulus information," in *Proc. IEEE Mil. Commun. (MILCOM)*, vol. 3, 1997, pp. 1372–1376.

[34] K. Alex, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, 2017, pp. 84–90.

[35] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[37] D. Li, R. Yang, X. Li, and S. Zhu, "Radar signal modulation recognition based on deep joint learning," *IEEE Access*, vol. 8, pp. 48515–48528, 2020.

[38] J. Guo, H. Zhang, J. Xu, and Z. Chen, "Pattern recognition of wireless modulation signals based on deep learning," in *Proc. IEEE 6th Int. Symp. Electromagn. Compat. (ISEMC)*, 2019, pp. 1–5.

[39] 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.

[40] J. Ma, S.-C. Lin, H. Gao, and T. Qiu, "Automatic modulation classification under non-Gaussian noise: A deep residual learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[41] P. Qi, X. Zhou, S. Zheng, and Z. Li, "Automatic modulation classification based on deep residual networks with multimodal information," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 21–33, Mar. 2021.

[42] C. Hou, G. Liu, Q. Tian, Z. Zhou, L. Hua, and Y. Lin, "Multisignal modulation classification using sliding window detection and complex convolutional network in frequency domain," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 19438–19449, Oct. 2022.

[43] Y. Lin, Y. Tu, Z. Dou, L. Chen, and S. Mao, "Contour Stella image and deep learning for signal recognition in the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 34–46, Mar. 2021.

[44] Z. He et al., "Deep learning-based automatic modulation recognition algorithm in non-cooperative communication systems," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2019, pp. 1–6.

[45] Y. Wu, X. Li, and J. Fang, "A deep learning approach for modulation recognition via exploiting temporal correlations," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2018, pp. 1–5.

[46] X. Zha, X. Qin, Y. Zhou, and H. Peng, "Power of deep learning for amplitude-phase signal modulation recognition," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, 2019, pp. 454–458.

[47] J. Xu, C. Luo, G. Parr, and Y. Luo, "A spatiotemporal multi-channel learning framework for automatic modulation recognition," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1629–1632, Oct. 2020.

[48] Z. Zhang, C. Wang, C. Gan, S. Sun, and M. Wang, "Automatic modulation classification using convolutional neural network with features fusion of SPWVD and BJD," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 5, no. 3, pp. 469–478, Sep. 2019.

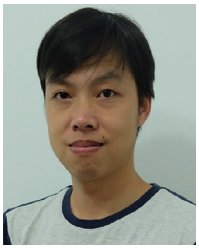
[49] Y. Shi, H. Xu, L. Jiang, and Y. Liu, "Few-shot modulation classification method based on feature dimension reduction and pseudo-label training," *IEEE Access*, vol. 8, pp. 140411–140425, 2020.

[50] X. Li, Q. Gong, S. Zhong, and S. Ren, "Near-field noncircular source localization based on fourth-order cumulant," *IEEE Access*, vol. 8, pp. 120575–120585, 2020.

[51] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[52] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," 2017, *arXiv:1707.01083*.

[53] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016, *arXiv:1610.02357*.



Shilian Zheng received the B.S. degree in telecommunication engineering and the M.S. degree in signal and information processing from Hangzhou Dianzi University, Hangzhou, China, in 2005 and 2008, respectively, and the Ph.D. degree in communication and information system from Xidian University, Xi'an, China, in 2014.

He is currently an Associate Researcher with the Science and Technology on Communication Information Security Control Laboratory, Jiaxing, China, and the Director of Zhejiang Signal Processing Society, Hangzhou. His research interests include cognitive radio, spectrum management, and deep learning-based radio signal processing.

Xiaoyu Zhou (Member, IEEE) was born in Anhui Province, China, in 1995. He received the B.S. degree in telecommunication engineering from Nanchang University, Nanchang, China, in 2019. He is currently pursuing the Ph.D. degree with Xidian University, Xi'an, China.

His research interests include signal processing and deep learning.

Luxin Zhang received the M.S. degree in control science and engineering from the Zhejiang University of Technology, Hangzhou, China, in 2021.

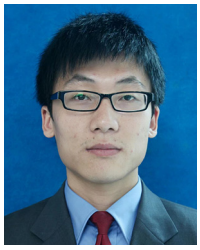
He is currently an Assistant Engineer with the Science and Technology on Communication Information Security Control Laboratory, Jiaxing, China. His research interests include cognitive radio, radio signal processing, and learning-based radio signal recognition.

Peihan Qi (Member, IEEE) was born in Henan, China, in 1986. He received the B.S. degree in telecommunications engineering from Chang'an University, Xi'an, China, in 2006, and the M.S. and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, in 2011 and 2014, respectively.

He has been a Professor with the School of Telecommunications Engineering, Xidian University since July 2022. His research interests include compressed sensing, spectrum sensing in cognitive radio networks, and intelligent signal processing.

Kunfeng Qiu was born in Shaoxing, Zhejiang, China, in 1997. He received the M.S. degree in control engineering from the Zhejiang University of Technology, Hangzhou, China, in 2022.

He is currently an Assistant Engineer with the Science and Technology on Communication Information Security Control Laboratory, Jiaxing, China. His research interests include radio signal processing and radio signal recognition based on deep learning.



Jiawei Zhu received the master's degree in electronic circuits and systems from the University of Electronic Science and Technology of China, Chengdu, China, in 2012.

He is currently an Associate Researcher with the Science and Technology on Communication Information Security Control Laboratory, Jiaxing, China. His research interests include radio signal processing, analysis and recognition, big data for radio signals, and deep learning algorithms for radio signals.

Xiaoniu Yang is currently a Chief Scientist with the Science and Technology on Communication Information Security Control Laboratory, Jiaxing, China. He published the first software radio book in China [X. Yang, C. Lou, and J. Xu, Software Radio Principles and Applications, Publishing House of Electronics Industry, 2001 (in Chinese)]. He holds more than 40 patents. His current research interests are software-defined satellite, big data for radio signals, and deep learning based signal processing. He is also an Academician of Chinese Academy of Engineering and a Fellow of the Chinese Institute of Electronics.