frontiers | Frontiers in Communications and Networks

TYPE Original Research
PUBLISHED 05 March 2025
DOI 10.3389/frcmn.2025.1453125

OPEN ACCESS

EDITED BY
Ramoni Adeogun,
Aalborg University, Denmark

REVIEWED BY
Man Fai Leung,
Anglia Ruskin University, United Kingdom
Khawla Alnajjar,
University of Sharjah, United Arab Emirates

*CORRESPONDENCE
Min Zhang,
minzhang@xidian.edu.cn

†These authors have contributed equally to this work

RECEIVED 13 August 2024
ACCEPTED 10 February 2025
PUBLISHED 05 March 2025

CITATION
Zhang H, Kuang Y, Huang R, Lin S, Dong Y and Zhang M (2025) Modulation recognition method based on multimodal features. *Front. Comms. Net* 6:1453125. doi: 10.3389/frcmn.2025.1453125

Modulation recognition method based on multimodal features

Hu Zhang†, Yin Kuang†, Ronghui Huang, Sheng Lin, Youqiang Dong and Min Zhang*

School of Aerospace Science and Technology, Xidian University, Xi'an, China

Introduction: Automatic modulation recognition (AMR) plays a crucial role in modern communication systems for efficient signal processing and monitoring. However, existing modulation recognition methods often lack comprehensive feature extraction and suffer from recognition inaccuracies.

Methods: To overcome these challenges, we present a multi-task modulation recognition approach leveraging multimodal features. In this method, a network is proposed to differentiate between multi-domain features for temporal feature extraction. Simultaneously, a network capable of extracting features at multiple scales is utilized for image feature extraction. Subsequently, recognition is conducted by integrating the multimodal features. Due to the inherent differences between 1D signal features and 2D image features, recognizing them collectively may overlook the unique characteristics of each type.

Results: We examine the merit of the proposed multi-task modulation recognition method and validate their performance with experiments using a public datasets. With an SNR of 0 dB, the proposed algorithm achieves a recognition accuracy of 92.30% on the RadioML2016.10a dataset.

Discussion: Therefore, we propose a multi-task modulation recognition approach leveraging multimodal features to enhance accuracy. By integrating temporal and image-based feature extraction, our method outperforms existing techniques in recognition performance.

KEYWORDS

automatic modulation recognition, feature extraction, multi-domain, multi-task, deep learning

1 Introduction

Automatic modulation recognition (AMR) plays a crucial role in modern communication systems, involving the automatic recognition of the received signal's modulation type, ensuring accurate decoding and understanding of information.

In the military domain (Ansari et al., 2022), AMR stands as a pivotal technology within electronic countermeasures, holding significant importance for strategic decision-making. In the civilian sector (Jdid et al., 2021; Hu et al., 2023; KHAN et al., 2017), AMR plays a critical role in spectrum resource monitoring. However, with the rapid advancement of communication technology, modulation methods have become increasingly diverse and complex, posing challenges to modulation identification. Therefore, studying the modulation identification of communication signals holds great theoretical and practical significance.

Traditional automatic modulation recognition (AMR) methods can be broadly categorized into two main types: those based on likelihood ratio tests and those relying on manual feature extraction. AMR methods based on likelihood ratio tests (zheng and Lv, 2018) are primarily

utilized to identify the modulation mode of a signal by constructing a likelihood function. Although effective, the high computational complexity of the method limits its application in increasingly complex communication environments. On the other hand, AMR methods based on manual feature extraction (Simic et al., 2021) determine the modulation type of a signal by comparing the extracted signal features with a preset threshold. However, this approach is time-consuming and labor-intensive, and its recognition performance heavily depends on the expertise of the operators. Moreover, when the communication environment changes, the recognition method based on the original features may not be adaptable, leading to significant degradation in recognition performance.

With the rapid advancement of deep learning technology, its applications have extended beyond image detection (Lin et al., 2023; Lin et al., 2022), opening up new research methods and opportunities in the field of modulation recognition.

O'Shea et al. (2016) were the first to propose the application of convolutional neural networks (CNN) for modulated signal recognition. This method simplifies the entire process compared to traditional identification methods. However, experimental results show that its maximum recognition accuracy is 74%, which still needs to be improved. Then, Daldal et al. (2019) used a Long Short-Term Memory (LSTM) model to identify six modulation signals. It successfully demonstrates that LSTM outperforms CNN-based recognition methods in recognizing signal timing modes. To address the increasingly complex signal types and improve the robustness of the algorithm, Li et al. (2023) devised a network architecture that combines Residual Networks with next iteration (ResNeXt) and Gated Recurrent Unit (GRU). ResNeXt network captures unique semantic features, while the GRU focuses on extracting temporal features. In order to recognize the complementary atrributes of these different features, they proposed a discriminative correlation analysis model. Simulation results demonstrate the superiority of this approach and provide a solid foundation for future feature analysis. Moreover, these findings promote the future application of feature fusion in AMR. Subsequently, researchers have explored the image modality of the signal. Daldal et al. (2020) achieved successful recognition of signal image modality by converting a one-dimensional signal into a two-dimensional time-frequency map through short-time Fourier transform and feeding it into CNN for recognition. This experimental result also presents a novel idea for subsequent modal recognition research.

In addition, many studies have explored the feasibility of multi-feature fusion methods in the field of AMR. Zhang et al. (2022) proposed R&CNN for underwater acoustic signal modulation recognition, which combines the automatic feature extraction and learning capabilities of recurrent neural network (RNN) and convolutional neural network (CNN) without manual feature extraction, and has the advantages of high precision and fast processing time. Huang et al. (2022) proposed OAE-EEKNN, an efficient recognition method based on optimized autoencoders and evaluation-enhanced K-nearest neighbor algorithms, which enables fast and high-precision identification of multiple modulation types in underwater acoustic channels. Wang et al. (2024) proposed One2ThreeNet, a method that rationalized underwater acoustic signals into time series, used One2Three blocks to extract signal time features from three microscales, and combined with two-current compression excitation (SE) block spatial feature extractor to synthesize and extract higher-level AMR spatial features for classification.Finally, some multi-feature fusion methods have been proved effective in other fields. For example, Li et al. (2024) proposed CurriFusFormer, which integrates course learning with a multi-feature fusion transformer model to deal with various patterns and ratios of lost data. Use spatial, temporal, and static features to generate accurate real-time estimates of missing values in different scenarios.

Overall, the AMR method utilizing deep learning enhances the performance of signal recognition. However, most deep-learning approaches to modulation recognition tend to utilize only a single modal feature of the signal. Additionally, due to the varying degrees of influence that different modal features have on recognition, these methods may not fully capture the signal's characteristics. To this end, we propose a multi-task modulation recognition method based on multi-modal features, referred to as MTL. MTL effectively achieves more accurate recognition of modulated signals. Specifically, this paper selects the Markov Transition Field (MTF) image and the original 1D signal for multi-mode feature fusion. Markov transition field method can transform 1D signal recognition task into image recognition task. At present, most automatic modulation recognition methods based on deep learning only use single-mode features without considering the complementarity and difference between multi-mode features. MTF images can reveal hidden patterns and structures in the 1D signal that may not be apparent in a one-dimensional representation of the 1D signal, and *vice versa*. For example, the 1D signal may contain immediate information about the time series, while the MTF encoding provides statistical information about the signal state transition; While the 1D signal emphasizes local features of the signal, including instantaneous amplitude, frequency, and phase changes, MTF coding emphasizes global features of the signal, including long-term dependencies between states and transition patterns. Moreover, according to the information entropy theory, the fusion of multi-modal features increases the information entropy of the system, and more signal details can be captured by sharing and integrating information among different modalities. At the same time, multimodal features can cover a larger feature space and improve the discriminative power of the classifier. In summary, this paper provides the following contributions.

- To differentiate the importance of various domain features in recognition, we devised a time-series model capable of assigning weights to multi-domain features. Specifically, the extracted multi-domain features are aggregated both horizontally and vertically. The probability of each feature in the horizontal direction is obtained using softmax, which serves as a weight coefficient. This coefficient is then multiplied by the features composed in the vertical direction to derive the final target feature.
- To enrich the features of the signal, we utilize Markov Transition Field (MTF) coding to convert the one-dimensional signal into an image. Subsequently, image features are extracted at multiple scales using the MTF model based on the focal modulation network. This model can extract local features while preserving global features, making the features more informative.

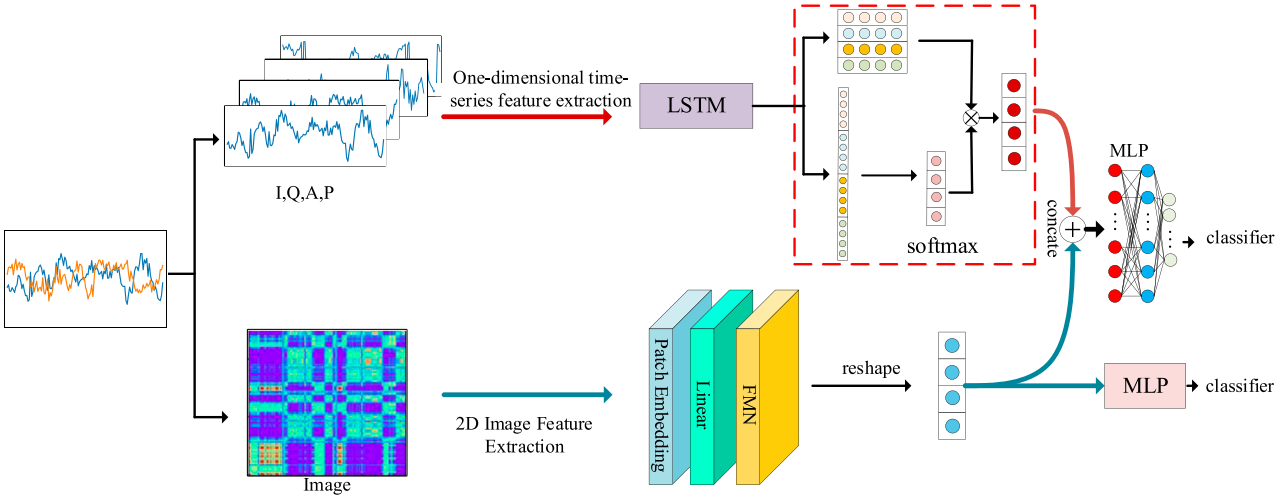

FIGURE 1
Diagram of the overall structure of the MTL model.

- To tackle the variability among different modalities and achieve better signal recognition, we propose a multi-task optimized recognition method. The main task is recognizing multimodal combinations, while the auxiliary task is recognizing the image modality. Different loss functions and weight coefficients are set for each task to enhance the recognition accuracy of the main task.

2 Signal model

In wireless communication systems, the modulated signals are transmitted by the transmitter, which are received by the receiver after propagation through the channel environment. The received signal $r(t)$ after downsampling at the t-th time slot can be expressed by Equation 1 as follows:

$$\begin{aligned} r(t) &= m(t) \cdot h(t) \cdot g(t) \cdot e^{j\left(2\pi\Delta f t+\theta\right)} + n(t) \\ t &= 0, 1, \ldots N-1 \end{aligned} \tag{1}$$

where $m(t)$ represents the baseband signal to be transmitted, and $n(t)$ denotes the additive Gaussian white noise. $h(t)$ and $g(t)$ denote the channel gain and pulse shaping response, respectively. Δf represents the carrier frequency offset and θ denotes the phase deviation. The received baseband signal $r(t)$ in complex form is generally given as an IQ component, as shown in Equation 2:

$$\begin{aligned} I(t) &= Re(r(t)) \\ Q(t) &= Im(r(t)) \end{aligned} \tag{2}$$

where the $Re(r(t))$ and $Im(r(t))$ represent the real and imaginary part of the received signal $r(t)$, respectively. The instantaneous amplitude $A(t)$ and the instantaneous phase $P(t)$ can be defined by Equation 3 as follows:

$$\begin{aligned} A(t) &= \sqrt{Re(r(t))^2 + Im(r(t))^2} \\ P(t) &= arctan\left(\frac{Im(r(t))}{Re(r(t))}\right) \end{aligned} \tag{3}$$

3 Methods

In this section, we introduce a detailed multimodal feature (MFF) recognition method. The method structure comprises two main parts: temporal modality-based modulation recognition and image modality-based modulation recognition. Subsequently, a multi-task loss optimization method is designed based on multimodal recognition, aiming to further enhance the overall recognition performance of the model and ensure full utilization of the advantages of each modality. The overall structure of the method is shown in Figure 1, where FMN denotes Focal Modulation Nets (Yang et al., 2022), MLP is Multilayer Perceptron. The *reshape* represents dimensional transformation, and the *concate* denotes feature concatenate.

3.1 Multimodal-based modulation recognition method

3.1.1 Temporal modal-based feature extraction module

Considering that LSTM has demonstrated certain advantages in processing sequence data due to its simple network structure, and given that the one-dimensional signals in this study are sequence data, LSTM was chosen as the basic network for extracting time-series features.

The modulation recognition method based on temporal modes proposed in this paper primarily relies on the LSTM model with improvements. The method can effectively assign multi domain features, enabling the model to better recognize the signal.

Firstly, the amplitude and phase (A/P) of the received signal are obtained through data transformation, and I/Q and A/P are jointly used as inputs to the model. Secondly, the LSTM model extracts features from the multi-domain data. The features extracted through LSTM are connected in two directions, horizontal and vertical, respectively. Then, the feature vectors in the horizontal direction undergo softmax processing to obtain the probability of each feature

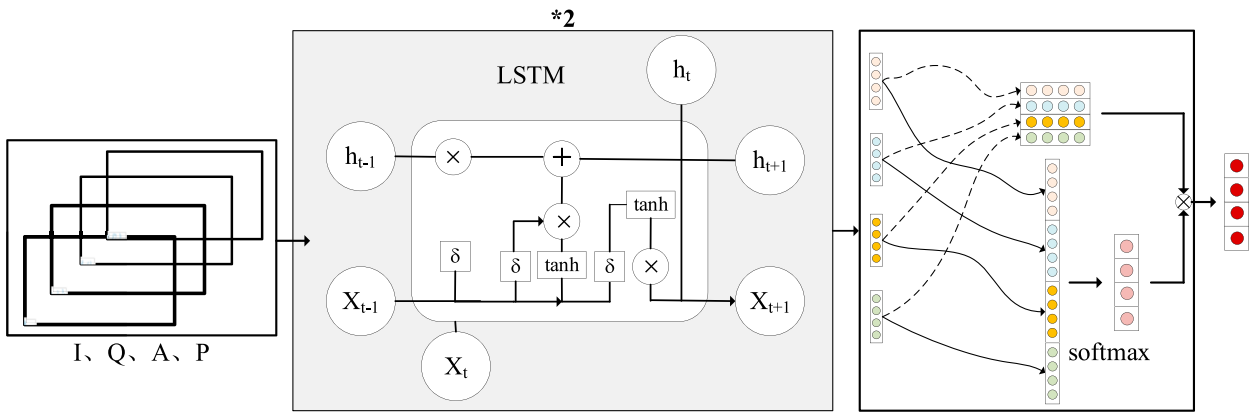

FIGURE 2
Schematic diagram depicting the structure of timing feature extraction.

$W = [w_1, w_2, w_3, w_4]$, which serves as a weighting coefficient. This coefficient is multiplied by the feature vectors in the vertical direction to obtain the final target features T_{final}. The calculation process can be described by Equation 4 as follows:

$$T_{final} = T{*}W^T = \begin{bmatrix} w_1 I_1 & w_1 I_2 & \cdots & w_1 I_n \\ w_2 Q_1 & w_2 Q_2 & \cdots & w_2 Q_n \\ w_3 A_1 & w_3 A_2 & \cdots & w_3 A_n \\ w_4 P_1 & w_4 P_2 & \cdots & w_4 P_n \end{bmatrix} \tag{4}$$

where T represents the eigenvector, composed by connecting in the vertical direction. The structure of the temporal model feature extraction is schematically depicted in Figure 2, where X_t and h_t denote input and hidden states at time step t, respectively. δ and tanh represent sigmoid activation function and hyperbolic tangent function.

3.1.2 Image modality-based feature extraction module

For the extraction of image modal features, the FMN network is chosen as the base network in this paper. FMN is capable of multi-scale feature extraction, enabling it to simultaneously capture global and local features. This capability allows FMN to retain richer and more effective features.

The image form not only increases the dimensionality of the data and provides more information to the model, but also helps to improve the expressive and generalization ability of the model. Image is the form of data representation in two-dimensional space, by converting time series into image, time series data can be combined with spatial information, so as to more comprehensively describe the characteristics of the data and the law of change. In this paper, the time series signal is converted into a two-dimensional image by using Markov transition field (MTF), an image coding method. This method can transform the one-dimensional signal recognition task into an image recognition task, extending the technical means of signal modulation recognition. At the same time, the MTF coding process also retains the dynamic statistical characteristics of the signal, which makes the information contained in it richer.

To realize one-dimensional signal imaging for a one-dimensional signal sequence, the following four steps are implemented:

Firstly, the sequence is divided into several segments, each labeled accordingly. Each segment contains the same number of sample points. Then, the weighted adjacency matrix is constructed by calculating the transition probabilities between the segments, which can be described by Equation 5 as follows:

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1Q} \\ w_{21} & \cdots & w_{2Q} \\ \vdots & \ddots & \vdots \\ w_{Q1} & \cdots & w_{QQ} \end{bmatrix}, \quad s.t. \sum_j w_{ij} = 1; i, j = 1, 2, ..., Q. \tag{5}$$

where w_{ij} denotes the transition probability from segment i to segment j, and the elements in each row are greater than 1 and the sum of all elements is equal to 1. Finally, the MTF matrix M is obtained by normalizing W and expressing the transfer probabilities in chronological order.

$$M = \begin{bmatrix} m_{ij} \,|\, x_1 \in q_i, x_1 \in q_j & \cdots & m_{ij} \,|\, x_1 \in q_i, x_N \in q_j \\ m_{ij} \,|\, x_2 \in q_i, x_1 \in q_j & \cdots & m_{ij} \,|\, x_2 \in q_i, x_N \in q_j \\ \vdots & \ddots & \vdots \\ m_{ij} \,|\, x_N \in q_i, x_1 \in q_j & \cdots & m_{ij} \,|\, x_N \in q_i, x_N \in q_j \end{bmatrix} \tag{6}$$

where q_i and q_j denote the ordinal numbers in the q-th quantile segment. The elements on the main diagonal represent the probability of self-transition. The MTF transition matrices M^I and M^Q for the I and Q signals can be derived from Equation 6, respectively. The MTF transition matrix M^{IQ} for the combined signals can be expressed by Equation 7 as follows:

$$M^{IQ} = M^I + M^Q \tag{7}$$

Afterwards, the image data is fed into the image feature extraction module for feature extraction. The structure of this module is depicted schematically in Figure 3, where H, W and d represent the height, width, and feature dimension of the feature map, respectively.

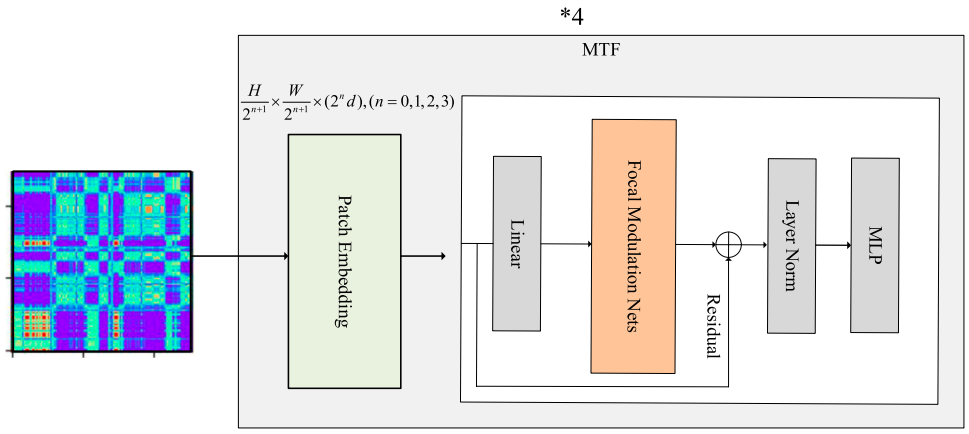

FIGURE 3
Schematic representation of the image feature extraction architecture.

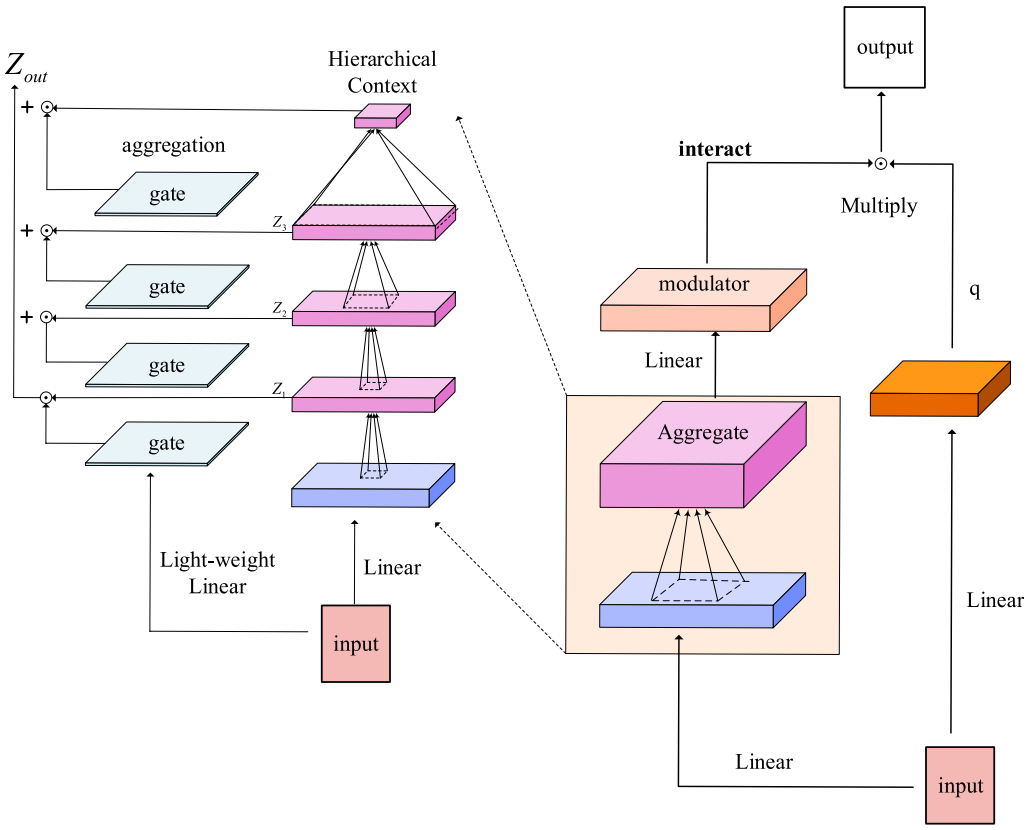

FIGURE 4
The structure of Focal Modulation Nets.

Initially, a one-dimensional signal is encoded into a two-dimensional image using MTF. Subsequently, a patch embedding layer, composed of convolutional blocks known as Patch Embedding, segments the feature map. This segmented map is then passed through the focus modulation network following a linear transformation. This entire process is repeated four times. After each iteration, the patch embedding layer reduces the spatial size of the feature map by half and doubles the feature dimension.

Focal modulation consists of three main components. The first is the hierarchical context. This component uses a series of deep

convolutional layers to encode visual context from near to far, allowing the model to capture local to global visual information at various levels of granularity. The second component is the gated aggregation mechanism. This mechanism selectively aggregates contexts from different levels of granularity based on the content of query markers, filtering out irrelevant information. The final component is elemental modulation. This part achieves the combination of context and query tokens through element-level multiplication operations, resulting in a refined representation.

The structure of the FMN is illustrated in Figure 4, where z_i denotes the i-th layer of contextual features. gate denotes a gated aggregation computation. ⊙ represents dot product. The input image is segmented into multiple tokens to form a feature map. Hierarchical contextualization is achieved by stacking deep convolutional layers, and generating contextual feature maps at various levels of granularity. The global context is captured using Global Average Pooling. Next, spatial and level-aware gating weights are generated via linear layers. These gating weights enable the model to control the extent of context aggregation from different granularity levels. Finally, a linear layer projects the input feature map into a new feature space. The modulators are then fused with the query tokens through element-level multiplication to produce the final feature representation.

3.2 Multi-task based modulation recognition method

Considering the disparities in the characteristics of various modalities, this section introduces a multi-task recognition approach that leverages multi-modal features for primary recognition and image feature recognition as a secondary method. Distinct loss functions are employed for each recognition task. Subsequently, these loss functions are weighted differentially, and their weighted sum is calculated to derive the ultimate objective loss function. The optimization of this loss function is iteratively refined to enhance the model's recognition capabilities. Consequently, the proposed methodology not only achieves the integration of multimodal features but also optimizes multiple loss functions, thereby augmenting the network's overall performance in sophisticated settings.

For recognizing multimodal features, we opted for the cross-entropy loss function, while for recognizing image modalities, we selected the KL divergence loss function. The formula for the objective loss function defined by Equation 8 is as follows

$$Loss_{\text{target}} = \alpha Loss_{Multi} + \beta Loss_{\text{Image}} \tag{8}$$

where α and β are the weight coefficients for two different recognition tasks. $Loss_{Multi}$ and $Loss_{\text{Image}}$ represent the loss functions for the multimodal feature recognition task and the image modal feature recognition task, respectively.

4 Experiments and results analysis

4.1 Experimental setup

4.1.1 Datasets

Experiments were conducted using the open-source dataset RadioML2016.10A (Lin et al., 2023), partitioned into training, validation, and test sets with a ratio of 6:2:2. The related parameters are illustrated in Table 1.

4.1.2 Evaluation indicators

In this paper, recognition accuracy and confusion matrix are selected as evaluation metrics. Under conditions of low SNR, the recognition accuracy of all models is significantly reduced. To clearly examine the advantages of different models, two representative SNR conditions, 0 dB and 18dB, are chosen for experimental analysis. The recognition accuracy can be expressed by Equation 9 as follows

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP and TN represent the number of correctly identified positive and negative samples, respectively, and FP and FN represent the number of incorrectly identified positive and negative samples, respectively.

The confusion matrix is a square matrix representing the total number of categories of the samples. The horizontal axis denotes the predicted results, while the vertical axis represents the true identification results. Each cell's value indicates the number of samples from the true category that are predicted to be in the corresponding predicted category. To provide a more intuitive representation of the confusion results, this study normalizes the number of predictions and then multiplies the values by 100, retaining two decimal places, resulting in a probability value.

4.1.3 Experimental setup

To determine the loss weight coefficients for the various tasks in the MTL model, a parametric analysis was conducted. The details of other specific parameters are presented in Table 2.

This paper conducts a comparative analysis of several mainstream deep learning-based methods, namely CNN (Tekbıyık et al., 2020), LSTM (Rajendran et al., 2018), ResNet (Liu et al., 2017), CLDNN (West and O'shea, 2017), and Transformer (Cai et al., 2022).

TABLE 1 Dataset related parameters.

Sample size	220000
Modulated signal type	11
SNR range	−20 dB:2 dB:18 dB
Signal length	128

TABLE 2 Parameters related to model training.

Optimizer	Adam
Batch size	64
Maximum training epochs	200
Initial learning rate	0.0001

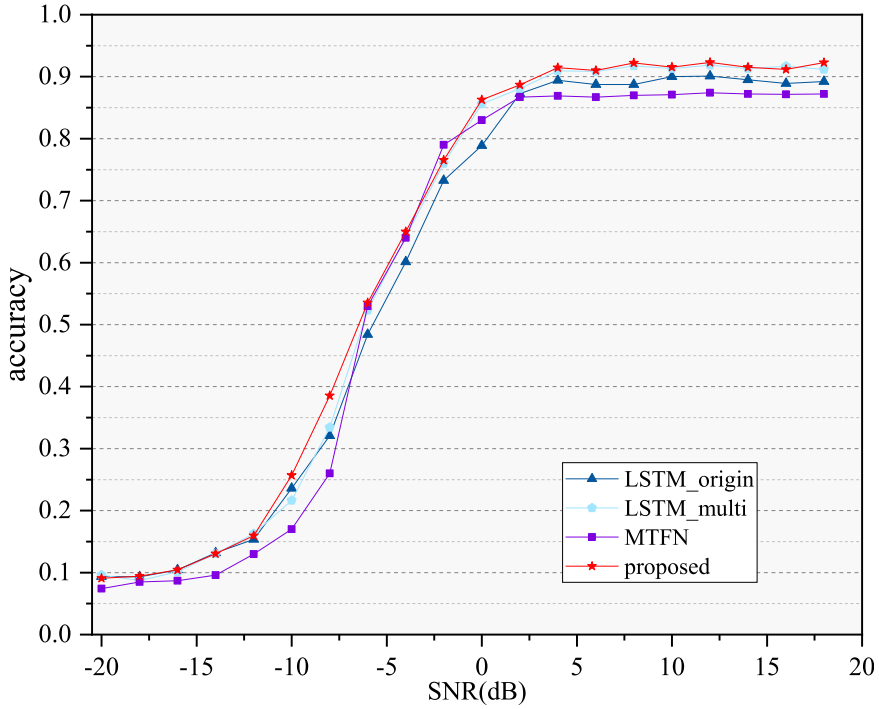

FIGURE 5
The recognition accuracy curves of MFF modules.

4.2 Performance of the proposed multimodal feature

4.2.1 Ablation study

We conducted ablation experiments on the modules in MFF and verified the effectiveness of each module. LSTM_orign denotes the original LSTM model that has not undergone any modification and.

LSTM_multi represents the improved LSTM model, which employs multiple features as input. MTFN stands for MTF Image Recognition Network. The proposed MFF is a multimodal recognition method. As illustrated in Figure 5, the LSTM_M method proposed in this study enhances recognition performance compared to the standard LSTM approach. This finding also confirms that features from different domains have varying impacts on the recognition outcomes. Furthermore, the recognition performance of the MFF method surpasses that of both LSTM_M and MTFN. This indicates that the richer effective features extracted by the multimodal recognition method contribute significantly to enhancing the model's recognition performance.

4.2.2 Comparison methods

In this paper, our proposed MFF method is compared with five mainstream modulation recognition methods. Figure 6 depicts the recognition accuracy curves of the six methods on the RML2016.10A dataset. When the SNR is below −12 dB, the recognition performance of all six methods degrades significantly due to the substantial impact of noise. However, when the SNR exceeds −10 dB, the proposed MFF model begins to demonstrate its recognition advantage for modulated signals. It can be found from the Table 3 that the best recognition accuracy of other methods is up to 90.10%, while the best recognition accuracy of the proposed MFF model reaches 92.30%. The average recognition accuracy is higher than CNN, ResNet, LSTM, CLDNN and Transformer. This also shows that the MFF is more stable in the recognition performance within 11 modulated signals, thereby ensuring a high average recognition accuracy.

To more intuitively observe the classification effect of the MFF method compared to several other models, Figure 7 presents the t-SNE visualization result plots for the six methods. As shown in Figure 7, the MFF method has the least confusing regions for classification, highlighting its recognition advantage over the other methods.

To further analyze the recognition performance of MFF, confusion matrices of the six methods at 0 dB are presented in Figure 8. Figures 8A–F depict CNN, LSTM, ResNet, CLDNN, Transformer, and the proposed method in this paper, respectively. As observed in Figure 8, at 0 dB, MFF has demonstrated effective recognition of two easily confused signals, 16QAM and 64QAM. However, for the WBFM signal, the presence of silence periods during sampling renders all six methods less effective in recognition.

Finally, since the proposed method employs two separate branch models to extract temporal and spatial features respectively, it is inevitable that the number of model parameters will expand. As illustrated in Table 4, while the proposed method achieves the optimal recognition performance, it involves a relatively larger number of parameters compared to other benchmark methods. This demonstrates that the improvement in recognition performance comes at the cost of a dramatic increase in the number of parameters. Importantly, the number of parameters remains within a manageable range, ensuring the model's practicality and feasibility in real-world scenarios. Specifically, the number of model parameters of the proposed method is 28.64 M, which is higher than that of CNN

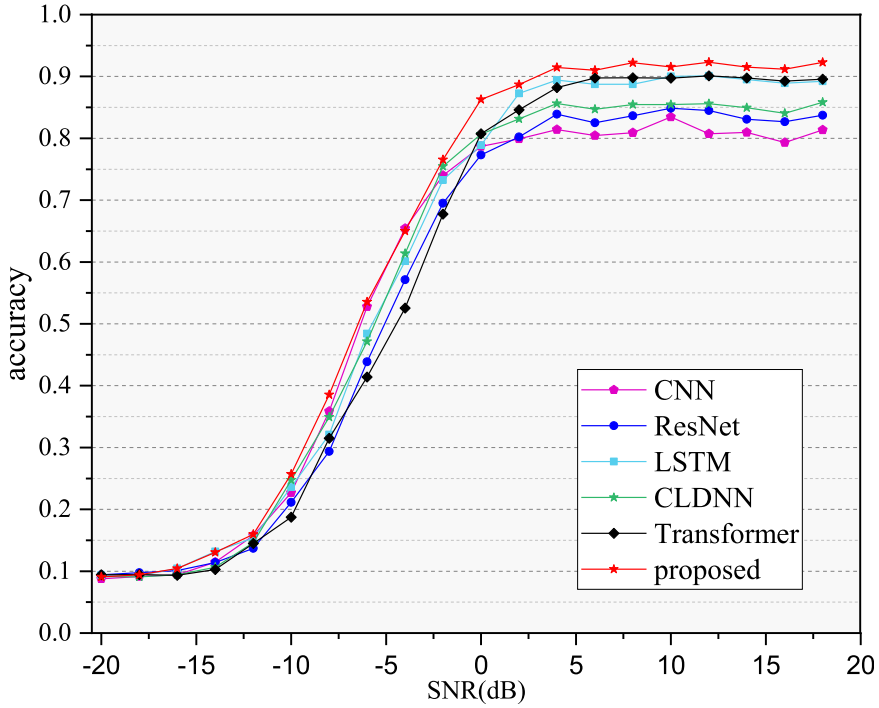

FIGURE 6
Recognition performance comparison of different methods.

TABLE 3 Comparative experimental results.

Model	Average accuracy	Best accuracy	0 dB	18 dB
CNN	55.62%	83.45%	78.68%	81.36%
ResNet	55.09%	84.86%	77.31%	83.72%
LSTM	58.97%	90.09%	82.50%	89.27%
CLDNN	57.09%	85.81%	80.68%	85.81%
Transformer	57.18%	90.10%	78.04%	89.57%
MFF	61.20%	92.30%	86.27%	92.27%

(1.59 M), ResNet (3.98 M), LSTM (0.26 M), CLDNN (0.51 M), and Transformer (11.53 M). This implies that the proposed algorithm has the lowest inference speed. While this design increases the computational requirements of the model, the method can achieve considerable performance gains and is suitable for scenarios with high accuracy requirements. With the continuous development of advanced devices with high computility, the MFF model will be a compelling example in the field of automatic modulation recognition, demonstrating that the modulation recognition performance can be significantly improved by increasing the model scale.

4.3 Performance of the proposed multi-task

4.3.1 Parametric analysis

To determine the coefficients for weighting loss in different tasks, we conducted experiments involving parametric analysis. The experimental results are depicted in Figure 9.

From Figure 9, it's evident that recognition accuracy increases initially and then decreases as α gradually decreases. Therefore, we selected the set of weight parameters that yielded the highest recognition accuracy, specifically $\alpha = 0.9, \beta = 0.1$. Across the four groups of experiments, the highest average recognition accuracy was nearly 1.5% higher than the lowest. With the increase of β from 0.05 to 0.1, all evaluation indexes of the model are rising. As the value of β increases from 0.1, the overall recognition performance and the best recognition performance exhibit a decreasing trend. The experimental result also further verifies that the multi-modal feature serves as the main recognition means and supplemented by the image mode, plays a positive role in the modulated signal recognition. Compared with other algorithms, this algorithm not only improves the recognition accuracy, but also enhances the ability of the signal characteristics description. Therefore, the algorithm combined with multi-modal and image modes can mitigate the effect of the complex wireless communication environment.

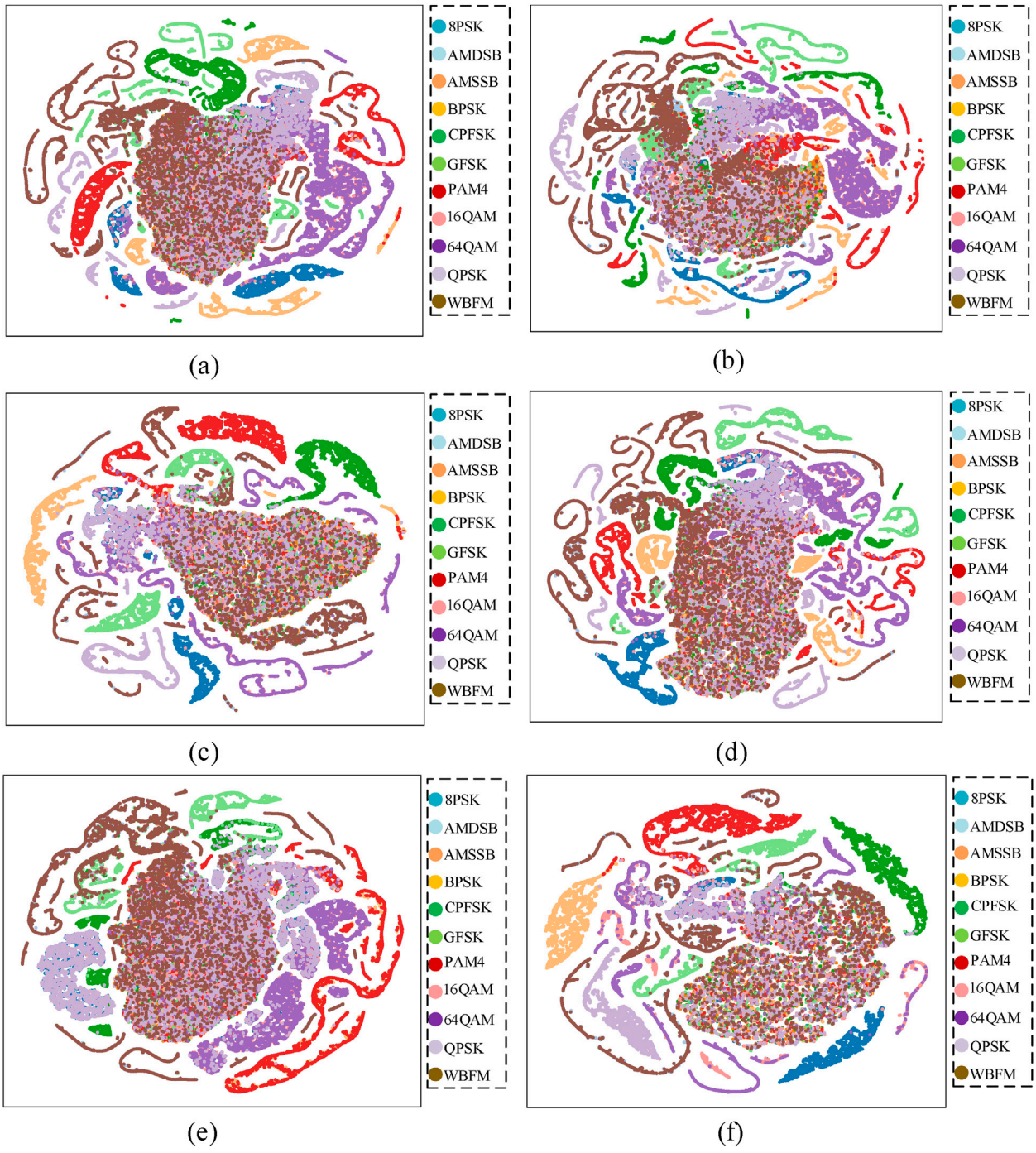

FIGURE 7
t-SNE visualization of classification effects of different models. **(A)** CNN. **(B)** ResNet. **(C)** LSTM. **(D)** CLDNN. **(E)** Transformer. **(F)** MFF.

4.3.2 Comparison experiment

To verify the superiority of the multi-task recognition method, additional experiments were conducted. The proposed MFF_ML model was compared with the MFF model, and the recognition accuracy variation curves for both models as functions of signal-to-noise ratio were obtained, as illustrated in Figure 10. The average and best recognition accuracy of MFF_ML were 62.33% and 92.90%, respectively. The recognition accuracy of 0 dB and 18 dB were 88.45% and 92.81%, respectively. Compared with the MFF method, the average recognition accuracy of MFF_ML is increased by 1.13%, and the best recognition accuracy is increased by 0.60%. It also verifies that the multi-task recognition method proposed in this chapter can further improve the recognition performance of multi-modal feature recognition.

The MFF_ML method proposed in this paper achieves the highest recognition accuracy of 92.9%, surpassing the five mainstream methods by at least 2.8%. The average recognition accuracy stands at 62.33%, underscoring the superiority of the MFF_ML method in modulated

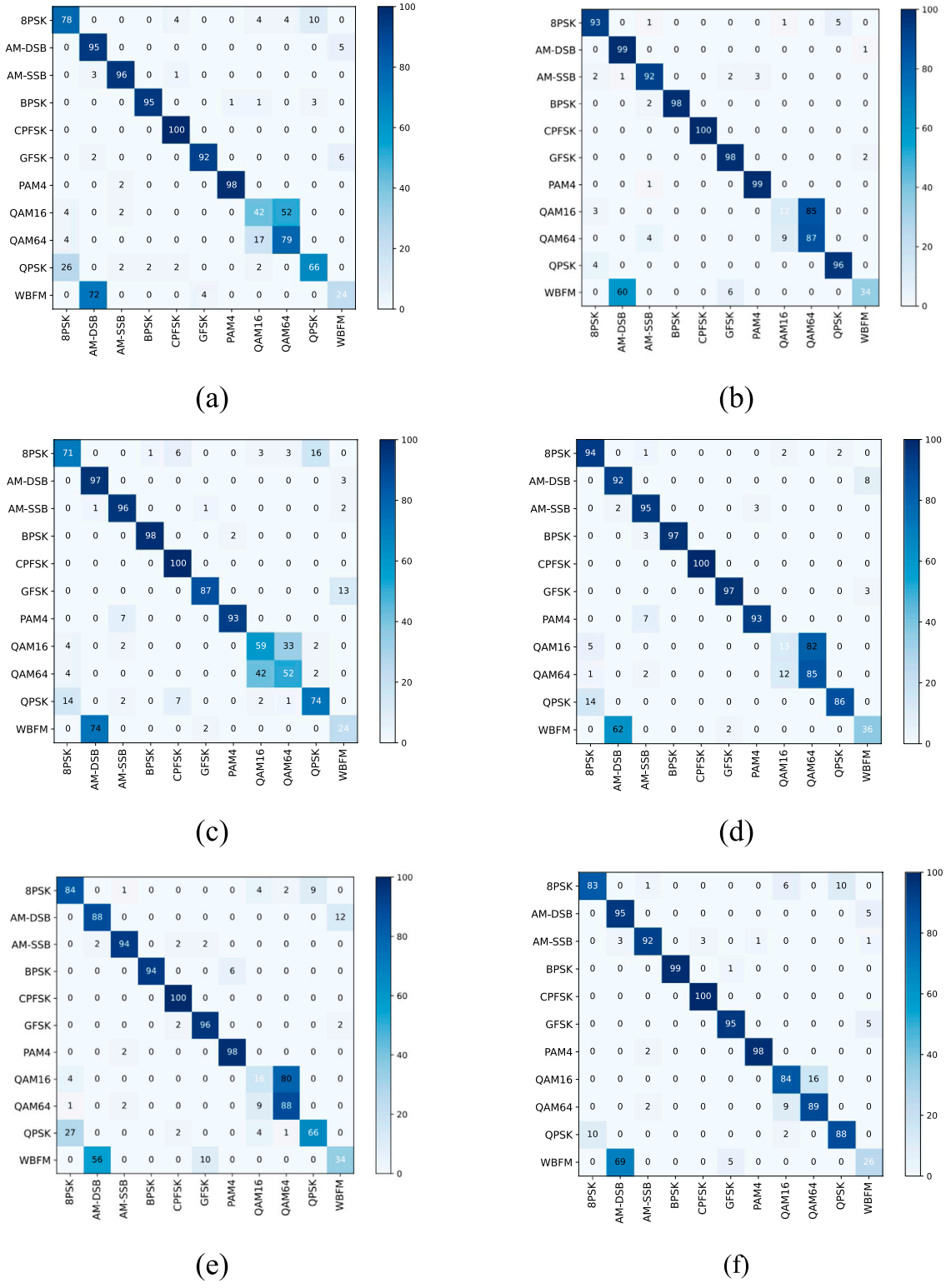

FIGURE 8
Confusion matrix between MFF and other methods at 0 dB. **(A)** CNN. **(B)** LSTM. **(C)** ResNet. **(D)** CLDNN. **(E)** Transformer. **(F)** MFF.

TABLE 4 The number of parameter comparison results.

Model	CNN	ResNet	LSTM	CLDNN	Transformer	MFF
Number of Parameters	1.59M	3.98M	0.26M	0.51M	11.53M	28.64M

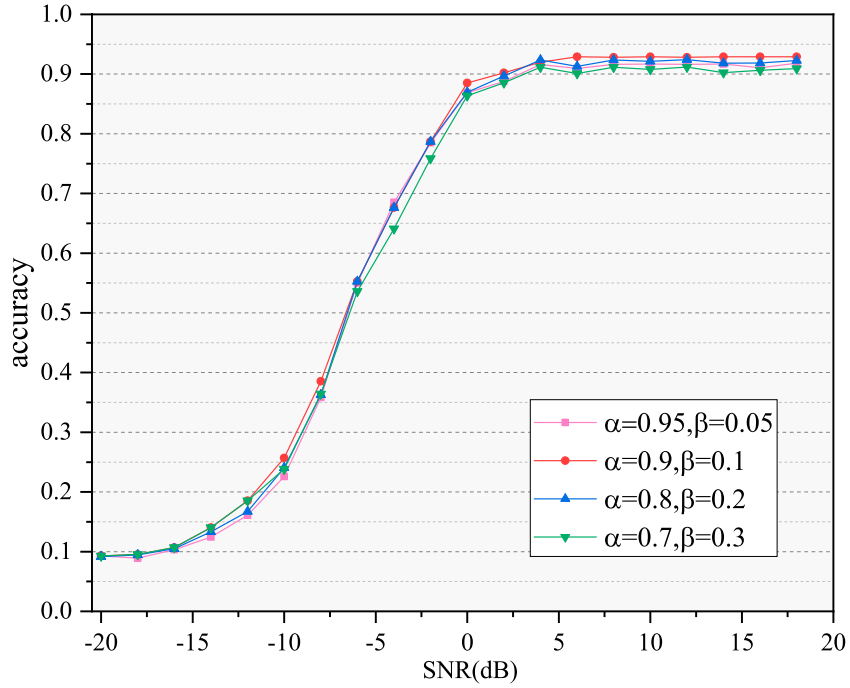

FIGURE 9
Recognition performance of MTL with different weighting coefficients.

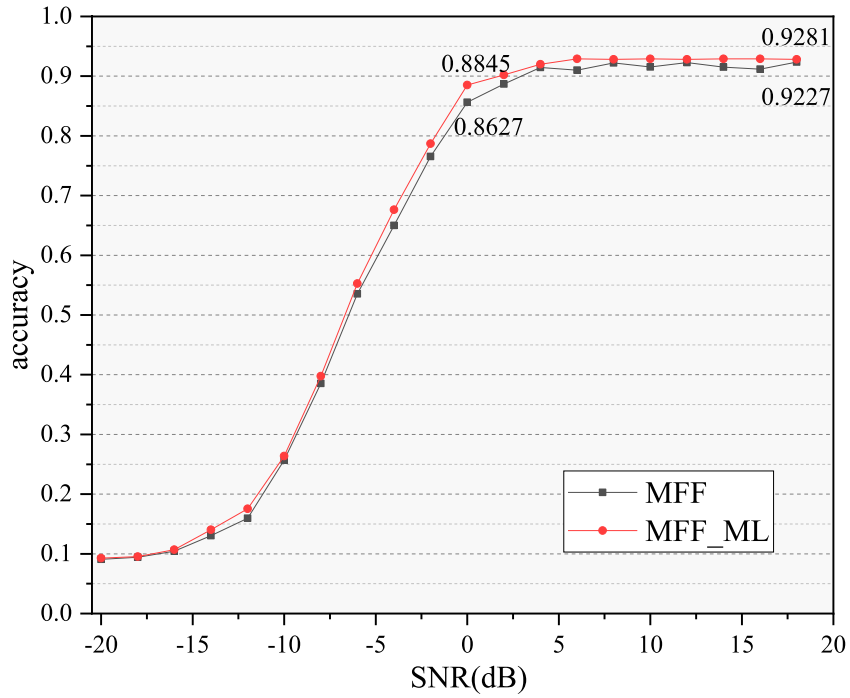

FIGURE 10
Recognition performance comparison of MFF_ML and MFF.

signal recognition. These findings demonstrate that the MFF_ML method offers superior and more stable recognition performance.

Figure 11 illustrates the confusion matrix at 0 dB for both methods.

As depicted in Figure 11, MFF_ML significantly enhances the discriminative capacity of 8PSK at 0 dB relative to MFF. This improvement can be attributed to the collaborative optimization of multiple tasks in MFF_ML, enabling the

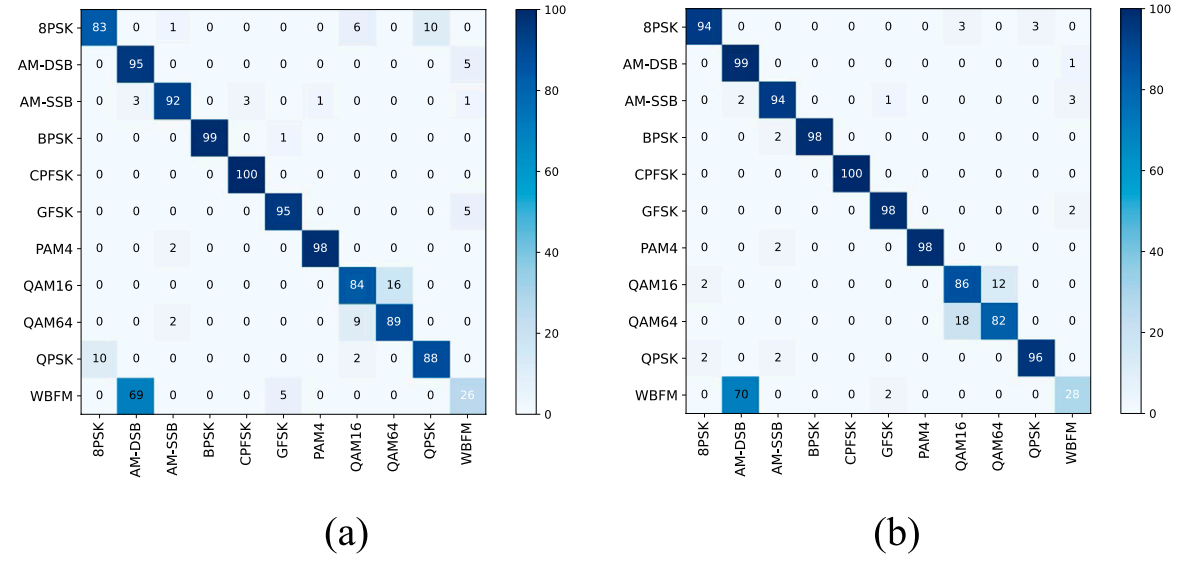

FIGURE 11
Confusion matrix between MFF and MFF_ML at 0 dB. **(A)** MFF. **(B)** MFF_ML.

model to extract more resilient and distinctive features from the input data.

5 Conclusion

In this paper, we propose a novel multi-task approach for modulation recognition, which significantly advances the field through several key contributions. First, we design a time-series feature extraction model that dynamically allocates weights to multi-domain features, optimizing their representation before extraction. Second, to enrich feature information, we employ MFF_ML coding to transform temporal IQ signals into image format, enabling the extraction of complementary spatial features. Third, we introduce a focus modulation network to comprehensively extract effective features from the image modality. Finally, we devise a multi-task recognition method to mitigate variability among multimodal features, further enhancing the model's recognition accuracy. Experimental results demonstrate that our proposed multi-task modulation recognition method achieves at least 2.8% higher accuracy compared to five mainstream modulation recognition methods. This improvement underscores the effectiveness of leveraging multi-modal features and highlights the distinct roles played by features from different modalities in the recognition process. The success of our approach not only validates the potential of multi-task learning in modulation recognition but also provides a robust framework for future research in signal processing and related fields. For future work, we identify several promising directions. First, incorporating denoising processes could enhance the model's performance at very low signal-to-noise ratios, addressing a critical challenge in real-world applications. Second, extending the proposed framework to other signal recognition tasks, such as speech or image recognition, could validate its generalizability and broaden its impact.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

HZ: Writing–original draft, Data curation. YK: Formal Analysis, Funding acquisition, Writing–review and editing. RH: Investigation, Methodology, Writing–original draft. SL: Software, Writing–review and editing. YD: Validation, Writing–original draft. MZ: Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (No. 12003018).

Acknowledgments

The authors would like to thank the reviewers for their valuableand detailed comments that are crucial in improving the quality of this paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ansari, S., Alnajjar, K. A., Saad, M., Abdallah, S., and El-Moursy, A. A. (2022). Automatic digital modulation recognition based on genetic-algorithm-optimized machine learning models. *IEEE Access* 10, 50265–50277. doi:10.1109/access.2022.3171909

Cai, J., Gan, F., Cao, X., and Liu, W. (2022). Signal modulation classification based on the transformer network. *IEEE Trans. Cognitive Commun. Netw.* 8(3), 1348–1357. doi:10.1109/tccn.2022.3176640

Daldal, N., Cömert, Z., and Polat, K.L (2020). Automatic Determination of digital modulation types with different noises using convolutional neural network based on time–frequency information. *Appl. Soft Comput.* 86, 105834. doi:10.1016/j.asoc.2019.105834

Daldal, N., Yildirim, Ö., and Polat, K. (2019). Deep long short-term memory networks based automatic recognition of six different digital modulation types under varying noise conditions. *Neural Comput. Appl.* 31, 1967–1981. doi:10.1007/s00521-019-04261-2

Hu, Y., Li, C., Wang, X., Liu, L., and Xu, Y. (2023). Modulation recognition of optical and electromagnetic waves based on transfer learning. *Optik* 291, 171359. doi:10.1016/j.ijleo.2023.171359

Huang, Z., Li, S., Yang, X., and Wang, J. (2022). OAE-EEKNN: an accurate and efficient automatic modulation recognition method for underwater acoustic signals. *IEEE Signal Process. Lett.* 29, 518–522. doi:10.1109/LSP.2022.3145329

Jdid, B., Hassan, K., Dayoub, I., Lim, W. H., and Mokayef, M. (2021). Machine learning based automatic modulation recognition for wireless communications: a comprehensive survey. *IEEE Access* 9, 57851–57873. doi:10.1109/access.2021.3071801

Khan, A. A., Rehmani, M. H., and Rachedi, A. (2017). Cognitive-Radio-based internet of things: applications, architectures, spectrum related functionalities, and future research directions. *IEEE Wirel. Commun.* 24 (3): 17–25. doi:10.1109/mwc.2017.1600404

Li, D., Tang, J., Zhou, B., Cao, P., Hu, J., Leung, M. F., et al. (2024). Toward resilient electric vehicle charging monitoring systems: curriculum guided multi-feature fusion transformer. *IEEE Trans. Intelligent Transp. Syst.* 25 (12), 21356–21366. doi:10.1109/TITS.2024.3456843

Li, L., Zhu, Y., and Zhu, Z. (2023). Automatic modulation classification using resnext-gru with deep feature fusion. *IEEE Transaction Instrum. Meas.* 72,1–10. doi:10.1109/tim.2023.3290301

Lin, S., Zhang, M., Cheng, X., Shi, L., Gamba, P., and Wang, H. (2023). Dynamic low-rank and sparse priors constrained deep autoencoders for hyperspectral anomaly detection. *IEEE Trans. Instrum. Meas.* 73, 1–18. doi:10.1109/tim.2023.3323997

Lin, S., Zhang, M., Cheng, X., Zhou, K., Zhao, S., and Wang, H. (2022). Hyperspectral anomaly detection via sparse representation and collaborative representation. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* 16, 946–961. doi:10.1109/jstars.2022.3229834

Liu, X., Yang, D., and El Gamal, A. (2017). "Deep neural network architectures for modulation classification," in 2017 51st Asilomar conference on signals, systems, and computers, Pacific Grove, CA, October 20–01 November 2017, 915–919. doi:10.1109/ACSSC.2017.8335483

O'Shea, T. J., Corgan, J., and Clancy, T. C. (2016). "Convolutional radio modulation recognition networks," in *Engineering applications of neural networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17*. Springer International Publishing, 213–226.

Rajendran, S., Meert, W., Giustiniano, D., Lenders, V., and Pollin, S. (2018). Deep learning models for wireless signal classification with distributed low-cost spectrum sensors. *IEEE Trans. Cognitive Commun. Netw.* 4 (3), 433–445. doi:10.1109/tccn.2018.2835460

Simic, M., Stanković, M., and Orlic, V. D. (2021). Automatic modulation classification of real signals in AWGN channel based on sixth-order cumulants. *Radioengineering* 30 (1), 204–214. doi:10.13164/re.2021.0204

Tekbıyık, K., Ekti, A. R., and Görçin, A., (2020). "Robust and fast automatic modulation classification with CNN under multipath fading channels," in 2020 IEEE 91st vehicular technology conference (VTC2020-Spring), Antwerp, Belgium, May 25–28 2020, 1–6. doi:10.1109/VTC2020-Spring48590.2020.9128408

Wang, J., Huang, Z., Shi, W., and Mao, S. (2024). One2ThreeNet: an automatic microscale-based modulation recognition method for underwater acoustic communication systems. *IEEE Trans. Wirel. Commun.* 23 (8), 10287–10300. doi:10.1109/TWC.2024.3371226

West, N. E., and O'shea, T. (2017). "Deep architectures for modulation recognition," in 2017 IEEE international symposium on dynamic spectrum access networks (DySPAN), Baltimore, MD, March 06–09, 2017, 1–6. doi:10.1109/DySPAN.2017.7920754

Yang, J., Li, C., and Dai, X. (2022). "Focal modulation networks," in 36th International Conference on Neural Information Processing Systems, New Orleans, LA, November 28–December 9, 2022, 4203–4217. doi:10.5555/3600270.3600574

Zhang, W., Yang, X., Leng, C., Wang, J., and Mao, S. (2022). Modulation recognition of underwater acoustic signals using deep hybrid neural networks. *IEEE Trans. Wirel. Commun.* 21 (8), 5977–5988. doi:10.1109/TWC.2022.3144608

Zheng, J., and Lv, Y. (2018). Likelihood-based automatic modulation classification in OFDM with index modulation[J]. *IEEE Trans. Veh. Technol.* 67 (9), 8192–8204. doi:10.1109/tvt.2018.2839735