

Enhancing Robustness in Automatic Modulation Classification via Energy-Guided Multi-Scale

Jiahua Zhou

jiahuazhou123@gmail.com

Liaoning Technical University <https://orcid.org/0009-0008-4051-7473>

Research Article

Keywords: Automatic Modulation Classification, Deep learning, RML22 dataset, Energy-Guided Temporal Saliency Attention

Posted Date: March 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9146515/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Enhancing Robustness in Automatic Modulation Classification via Energy-Guided Multi-Scale Temporal Networks

Jiahua Zhou^{1*}

^{1*}Liaoning Technical University, School of Electronic and Information Engineering, Longwan Street, Huludao, 125105, Liaoning, China.

Corresponding author(s). E-mail(s): 2306030226@stu.lntu.edu.cn;

Abstract

Automatic Modulation Classification (AMC) is pivotal in non-cooperative communication systems. This study addresses the challenges of feature coupling and noise-induced degradation, particularly under low signal-to-noise ratio (SNR) conditions. We introduce the Energy-Guided Multi-Scale Temporal Network (EMST-Net), which employs a Multi-Stream Fusion Convolution module to decouple in-phase and quadrature components, enhancing feature expressiveness. An Energy-Guided Temporal Saliency Attention mechanism adaptively emphasizes informative signal segments while suppressing noise. A gated dilated temporal convolutional network captures both short-term and long-range dependencies. Extensive experiments on three public datasets (RML22, RML2016.10a, and RML2018.01a) demonstrate that EMST-Net achieves overall accuracies of 71.59%, 62.79%, and 61.74%, respectively, outperforming comparative models. Notably, it exhibits strong robustness under low SNR conditions (e.g., 86.27% accuracy at 0 dB on RML22). The model is computationally efficient, with only 260K parameters, 24.2M FLOPs, and a storage overhead of 0.811MB. The source code is available at: <https://github.com/jiahua Zhou123-cpu/EMST-Net>.

Keywords: Automatic Modulation Classification, Deep learning, RML22 dataset, Energy-Guided Temporal Saliency Attention

1 Introduction

Automatic Modulation Classification (AMC), as a key intermediary technology between signal detection and demodulation, can autonomously identify the modulation scheme of received signals without prior knowledge of transmission parameters in wireless communication [1, 2]. It plays a critical role in non-cooperative scenarios, including civil spectrum monitoring, military electronic reconnaissance, and cognitive radio [3, 4].

Conventional AMC methods mainly separated into two paradigms: likelihood-based (LB) and feature-based (FB) frameworks. LB framework interprets modulation classification as a hypothesis-testing problem [5]. Although this framework can theoretically offer the best performance under ideal conditions, it has extremely high computational complexity, depends on prior information, and remains the vulnerable to channel interference. FB framework first extracts artificial features from raw signals and then feeds them into conventional machine learning classifiers [6]. Compared with LB framework, FB demands less prior information and exhibit lower computational complexity. However, this method requires labor-intensive feature engineering and intricate feature selection procedures.

In recent years, the success of deep learning (DL) in various areas has inspires its adoption in AMC, as DL can autonomously extract signal features from original datasets and address the limitation of manual feature design. However, since the actual physical channel is complex and full of noise, it is inevitable that it will have an impact on the signal. Although the current AMC frameworks based on DL has made significant progresses, it still faces several severe challenges which constrain their discriminative capability and robustness:

- a. Most frameworks process IQ signals as a two-channel real-valued tensor without explicitly separating the statistical properties of the I and Q components. This entangled representation hinders the extraction of fine-grained constellation and amplitude-phase features, particularly for high-order modulations.
- b. DL models robust to AWGN often struggle in non-stationary channels. Equal weighting of temporal segments ignores sparse high-SNR features, leading to suboptimal aggregation and degraded robustness.

To address the aforementioned issues, this paper proposed a novel Energy-Guided Multi-Scale Temporal Network (EMST-Net) for AMC. Specially, a multi-scale fusion convolution module is introduced to explicitly decouple and jointly model the coupling relationship between I and Q components. This design enables the network to learn more expressive and diverse signal representations from multiple streams. Building upon this representation, we introduced an energy-guided temporal saliency attention mechanism to adaptively highlight informative signal segments and enhance physically meaningful modulation patterns. This module combines temporal saliency modeling with global feature calibration to capture both local signal features and long-range temporal dependencies. Then a gated dilated temporal convolution network further expand the temporal receptive field while maintaining stable gradient propagation. Finally, a robust classifier is introduced to obtain the classification result of the modulated signal. Through the collaboration of these modules, the model achieves more robust and discriminative feature learning for AMC.

What’s more, the main content and contributions of this paper are summarized as follows:

- a. We introduce a Multi-Scale Fusion Convolution Module (MSFC) with a tri-stream architecture for IQ signal, explicitly decoupling the complex baseband signals to enhance representation disentanglement and feature expressiveness.
- b. We propose an Energy-guided Temporal Saliency Attention (ETSA) Module that serves as a learnable saliency gating mechanism, utilizing channel energy to recalibrate responses and effectively suppressing transient noise.
- c. Extensive experiments on three public benchmark datasets (RML2016.10a, RML2018.01a, and RML22) demonstrate that the proposed ESTM-Net model achieves state-of-the-art performance. Notably, ESTM-Net consistently outperforms comparison methods across all SNR levels and has the most appropriate computational complexity and storage overhead.
- d. The proposed framework is further assessed through comprehensive evaluations. Beyond ablation experiments that measure the contribution of each component, we analyze class-wise recognition accuracy and apply t-SNE visualization to explore the separability of the learned feature embeddings.

2 Related Work

2.1 Evolution of DL in AMC

The application of Deep Learning (DL) in Automatic Modulation Classification (AMC) has witnessed significant progress since the pioneering work of O’Shea et al. [7], who transformed original radio IQ signals into grid-like representations for Convolutional Neural Networks (CNN). Subsequently, various temporal models such as RNNs [8], LSTMs [9, 10], and CLDNNs [11, 12] have been widely adopted to capture sequential dependencies. To address performance under varying SNR conditions, recent efforts have introduced sophisticated architectures: Gao et al. [13] proposed a Mixture-of-Experts (MoE-AMC) model for balanced recognition, while Liao et al. [14] developed the SCRNN model to significantly boost accuracy in high SNR ranges. This trend of applying specialized sequence modeling is also echoed in broader visual computing; for instance, Singh [15] demonstrated that soft attention coupled with Bi-LSTM can improve dynamic pattern recognition, illustrating a cross-domain requirement for robust feature focus in complex sequences.

2.2 Multi-branch and Multi-scale Temporal Modeling

To address the feature coupling inherent in raw I/Q data, multi-stream architectures have become a major research focus. Xu et al. [16] utilized a three-stream framework to extract features from individual and combined I/Q components, while Cheng et al. [17] proposed CVCNN-LSTM to surpass existing SOTA performance. This multi-branch philosophy is conceptually aligned with recent advances in visual computing, such as SCNet [18], which employs a dual-branch network to handle strong noise. Furthermore, the importance of multi-scale context has been underscored by Xue et al. [19] in sign language recognition, where capturing features across various temporal scales is

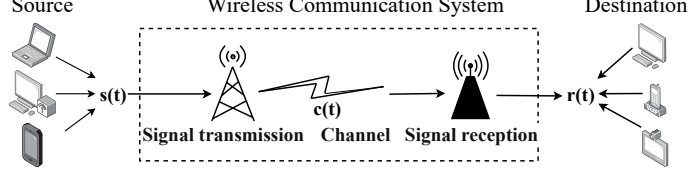


Fig. 1 Wireless communication system model

essential for understanding complex hierarchies. Similarly, Songchen Xu et al. [20] proposed the MCCSAN framework with innovations tailored for multi-channel AMC. Our proposed EMST-Net integrates gated dilated convolutions to expand the receptive field, allowing for the capturing of multi-scale temporal features while maintaining computational efficiency.

2.3 Attention Mechanisms and Robust Representation

Attention mechanisms have revolutionized the way networks prioritize informative segments while suppressing noise. Zhang et al. [21] introduced the WNRST model with weight non-negative constraints to optimize feature focus. Gu et al. [3] proposed LMDF-Net, a lightweight network leveraging channel-grouped self-attention to reduce parameters by 60%. More recently, Wu et al. [22] integrated vision transformers with Kolmogorov-Arnold Networks (KAN) in their MSFF-ViT model.

A particularly innovative direction is "energy-guided" learning. Pei et al. [23] introduced an energy-guided adaptation method for multi-modal perception, providing a theoretical foundation for using energy metrics. Similarly, structure-aware attention in SAT-Net [24] and artifact-consistent networks like CCM-Net [25] demonstrate that adaptively emphasizing structural integrity can significantly boost performance in low-quality data. These insights, combined with masked transformer frameworks [26] and unsupervised contrastive learning [27], underscore the necessity of adaptive saliency mechanisms—such as our Energy-Guided Temporal Saliency Attention—in achieving robust signal classification.

3 Signal Model

Wireless communication system is mainly composed of a transmitter, a channel and a receiver, as shown in Figure 1. However, the physical channel is complex and full of noise, the signal through the channel will be affected by noise inevitably. Therefore, the received signal should be represented as [14, 28]:

$$r(t) = s(t) * h(t) + n(t) \quad (1)$$

where $s(t)$ denotes the transmitted signal, $h(t)$ denotes the channel impulse response, the operator $*$ denotes the convolution sum operation, and $n(t)$ denotes the additive noise.

In recent years, with the rapid increase of communication services, In-phase and Quadrature Modulation (IQ modulation) as an efficient signal modulation technology receives widespread application. In IQ modulation, the input signal is first decomposed

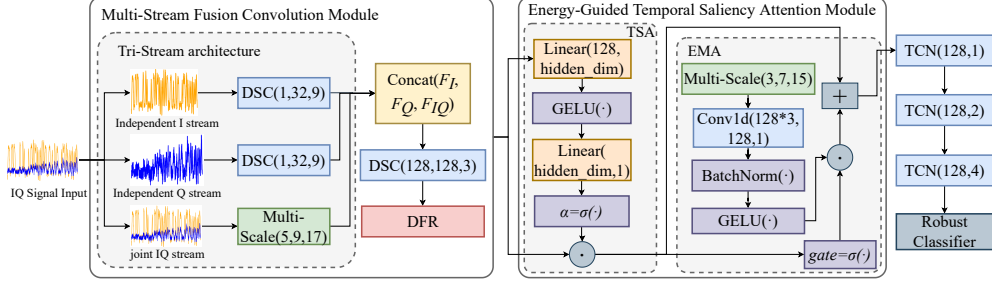


Fig. 2 The framework of EMST-Net for AMC

into in-phase (I) and quadrature (Q) components [29]. Following modulation onto synchronous sine/cosine carriers, the combined IQ signal is transmitted. Therefore, the received signal can be updated as [17, 28]:

$$r(t) = i(t) + j \cdot q(t) \quad (2)$$

where $i(t)$ denotes the in-phase component of the received signal and $q(t)$ denotes the quadrature component.

4 Method

This section introduces the proposed Energy-Guided Multi-Scale Temporal Network (EMST-Net). The overall architecture is designed to improve the robustness and discriminative capability of DL-based AMC under complex wireless environments. As shown in Figure 2, The framework consists of three major components: the Multi-Stream Fusion Convolution module (MSFC), the Energy-Guided Temporal Saliency Attention module (ETSA), and gated dilated Temporal Convolutional Networks (TCN) followed by a robust classifier.

The MSFC performs structured multi-stream feature extraction to explicitly model the relationships between IQ, I, and Q representations. Building upon these representations, the ETSA module enhances informative signal structures through temporal saliency estimation and energy-aware multi-scale aggregation mechanism. The extracted features are then processed by dilated temporal convolution blocks to capture long-range dependencies across symbol sequences. Finally, a robust classifier produces the modulation schemes predictions.

4.1 Multi-Stream Fusion Convolution Module

In this section, we introduce the proposed Multi-Stream Fusion Convolution module (MSFC), which is designed to enhance the temporal representation capability of complex baseband signals. This module adopts a tri-stream architecture, a multi-scale computing to enhance the feature extraction capability of the convolutional layer and a Dual-Domain Feature Recalibration module (DFR) to integrate the global channel statistical information with the local temporal. Through this structured design, the module aims to achieve effective feature fusion.

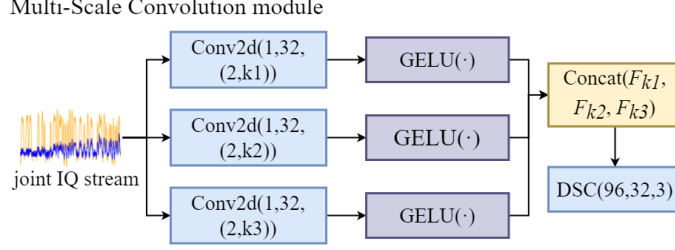


Fig. 3 Multi-Scale Convolution module

4.1.1 Tri-Stream architecture and Multi-Scale Convolution module

Conventional AMC methods based on DL typically treat the received IQ signal as a two-channel real-valued input and process through a single convolutional pathway for joint feature extraction. Although this strategy can learn global modulation patterns, the I and Q components are projected into same feature space and implicitly fused during convolutional operations without explicit structural separation. This may lead to entangled representations, where inter-component coupling and component-specific statistical characteristics are implicitly fused, limiting the efficiency of discriminative feature modeling. To solve this issue, we construct a tri-stream architecture: a joint IQ stream for modeling inter-component coupling and two independent streams for capturing I and Q-specific statistical dynamics, thereby enhancing feature expressiveness and representation disentanglement [16, 29].

In the tri-stream architecture, the joint IQ stream adopts multi-scale convolution, whereas the I and Q streams employ single-scale convolution. This different design is motivated that the joint IQ stream contains multi-scale amplitude/phase interactions and constellation structures, requiring multi-scale convolution to capture hierarchical temporal patterns effectively. In contrast, the individual I and Q streams exhibit relatively localized statistical characteristics, for which single-scale convolution is sufficient and more computationally efficient. This module is shown in Figure 3.

Suppose the received signal is in the form of $X_{IQ} \in \mathbb{R}^{B \times 2 \times T}$, where B denotes the batch-size, T denotes the temporal length, and the second dimension corresponds to the I and Q components. The I and Q components are separated as $X_I \in \mathbb{R}^{B \times 1 \times T}$ and $X_Q \in \mathbb{R}^{B \times 1 \times T}$.

In the joint IQ stream, IQ signal is first expanded into a two-dimensional form which is reshaped as $X \in \mathbb{R}^{B \times 1 \times 2 \times T}$. A set of two-dimensional convolutional operators with kernel size $(2, k)$ is applied, where the kernel fully spans the I/Q dimension and slides along time. For a given temporal scale k , the 2D convolution produces a 32-dimension feature vector at each temporal position t , which can be defined as:

$$F^{(k)}(t) = \sum_{\tau}^{k-1} W^{(k)}(\tau) X_{IQ}(t - \tau) \quad (3)$$

where $X_{IQ}(t) \in \mathbb{R}^2$ denotes the joint IQ vector at time t , and $W^{(k)}(\tau) \in \mathbb{R}^{(C \times 2)}$ denotes the learnable projection matrix associated with temporal offset τ . Because the kernel spans both I and Q channels, this operation explicitly couples them at

each time step. Consequently, it projects local complex-valued signals into a higher-dimensional space, allowing the network to capture joint amplitude-phase variations and constellation geometry.

To extract features with multi-scale information, three branches of convolutional layers with different kernel size are organized in parallel [30]. The kernel sizes of these branches are $k_1=5, k_2=9, k_3=19$. These branches of the convolutional layers increase the width of the network and enable the model to learn multi-scale information [23]. Features from different convolutional layers are consolidated via a channel-wise concatenation, which enhances the model's capacity to characterize complex signal patterns:

$$F'_{IQ} = \text{Concat}(F_{IQ}^{(5)}, F_{IQ}^{(9)}, F_{IQ}^{(19)}) \quad (4)$$

where $F_{IQ} \in \mathbb{R}^{B \times 3C \times T}$. This concatenation operation preserves temporal alignment while integrating heterogeneous receptive-field responses into a unified representation. To further extract feature representation, F_{IQ} is subsequently processed by a 1D depth-wise separable convolution (DSC) with a small kernel ($k=3$) to achieve dimension compression. This is followed by a pointwise convolution to combine the outputs of the depth-wise convolution [31]:

$$F_{IQ} = \text{Dropout}(\text{GELU}(\text{BN}(W_p * W_d * F'_{IQ}))) \quad (5)$$

where W_d represents the depth-wise one-dimensional convolution, W_p represents the 1×1 pointwise convolution, $\text{BN}(\cdot)$ denotes the batch normalization operation, $\text{GELU}(\cdot)$ denotes the Gaussian Error Linear Unit activation function and $\text{Dropout}(\cdot)$ denotes dropout regularization operation. After regularization via dropout, the feature tensor rearranged into $\mathbb{R}^{B \times T \times D_{IQ}}$, making it suitable for subsequent modules.

Compared to the complex joint IQ stream, the I and Q streams extract signal features through the DSC with a high kernel ($K=9$):

$$F_I = \text{Dropout}(\text{GELU}(\text{BN}(W_p * W_d * F'_I))), X_i \in X_I, X_Q \quad (6)$$

where $F_i \in \mathbb{R}^{B \times T \times D}$. Subsequently, the feature tensor of the I and Q streams are combined with this of IQ stream to obtain the convolutional feature based on the three parallel stream:

$$F = \text{Concat}(F_{IQ}, F_I, F_Q) \quad (7)$$

where $F \in \mathbb{R}^{B \times T \times D}$ and $D = D_I + D_Q + D_{IQ}$.

4.1.2 Dual-Domain Feature Recalibration module

To further enhance discriminative capacity, a Dual-Domain Feature Recalibration (DFR) module is introduced. DFR module performs a global channel-wise recalibration followed by depth-wise temporal gating, thereby adaptively modulating the fused feature representation along both channel and time dimensions, which is shown in Figure 4.

Before applying the DFR module, a DSC is introduced to further refine and fuse the cross-stream features. Subsequently, F is inputted into DFR. First, we compress the

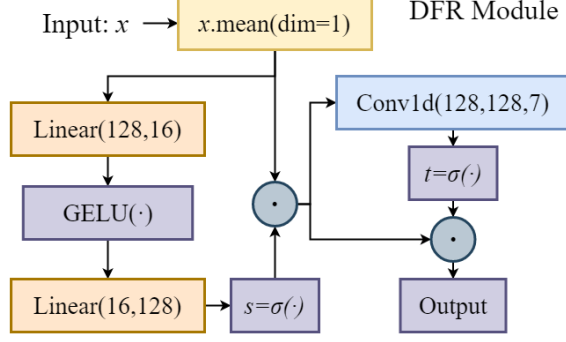


Fig. 4 DFR module

temporal redundancy by applying Global Average Pooling (GAP) across the sequence length T . This operation yields a global descriptor $S \in \mathbb{R}^{B \times D}$, which encapsulates the global distribution of each channel:

$$S = \frac{1}{T} \sum_{t=1}^T F_{:,t,:}$$

To learn the complex dependencies between channels, S is projected into a lower-dimensional latent space with a reduction ratio $r = 8$ and then re-expanded to the original channel count. This process is governed by a gating mechanism using the Sigmoid activation function, producing a scaling vector $S_{weight} \in \mathbb{R}^{B \times D}$:

$$S_{weight} = \sigma(W_{DFR2} \cdot GELU(W_{DFR1} \cdot S)) \quad (8)$$

where $W_{DFR2} \in \mathbb{R}^{\frac{D}{r} \times D}$ and $W_{DFR1} \in \mathbb{R}^{D \times \frac{D}{r}}$ represent learnable linear transformations, $\sigma(\cdot)$ denotes the GELU activation function (Sigmoid function). The intermediate channel-calibrated feature F_c is subsequently obtained by scaling the original input F with these weights:

$$F_c = F \odot S_{weight} \quad (9)$$

where $\odot$ denotes the Hadamard product via broadcasting along the temporal dimension. Then, F_c undergoes a 1-D convolution utilizing a kernel size $K = 7$ and padding $P = 3$, to extract the local temporal patterns of each channel, effectively distinguish the temporal dynamics from the channel-specific dynamics. The temporal recalibration mask $T_{weight} \in \mathbb{R}^{B \times D \times T}$ is represented as:

$$T_{weight} = \sigma(Conv1d(F_c^T))^T \quad (10)$$

Then the final output F_{out} is generated by scaling F_c with T_{weight} , followed by a Dropout layer to enhance robustness:

$$F_{out} = Dropout(F_c \odot T_{weight}) \quad (11)$$

The above descriptions present the overall design of MSFC. By integrating tri-stream architecture, multi-scale convolution module and DFR module, the proposed module effectively enhances feature discriminability and robustness in complex wireless environments.

4.2 Energy-Guided Temporal Saliency Attention Module

Due to the sparse distribution of discriminative features caused by time-varying channel fading, pulse interference, and corrupted or noisy modalities, AMC faces significant challenges in adapting to distribution shifts during testing [23].

Due to the sparse distribution of the discriminative features caused by time-varying channel fading and pulse interference and , AMC is faced with significant obstacles. To address this issue, we propose a novel architecture designed for simultaneous temporal saliency modeling and energy-weighted channel response extraction Energy-Guided Temporal Saliency Attention (ETSA). This module consists of temporal saliency attention (TSA) and energy-aware multi-Scale aggregation (EMA). As shown in Figure 5.

4.2.1 Temporal Saliency Attention mechanism

We introduced Temporal Saliency Attention (TSA) to address the issue that the non-stationary nature of wireless signals and the presence of pulse interference lead to the uneven distribution of discriminatory features over the time dimension. TSA can adaptively weight the significance of each time step, focus on the temporal evolution of the feature set, and prioritize the information signal segments while suppressing transient noise.

For each time step feature tensor $X \in \mathbb{R}^{B \times T \times C}$ and vector $x_t \in \mathbb{R}^C$. TSA generates a scalar saliency score for each time step through a lightweight two-layer multilayer perceptron (MLP). Specially, the saliency score is computed as:

$$s_t = W_{TSA2} \cdot GELU(W_{TSA1} \cdot x_t) \quad (12)$$

where $W_{TSA1} \in \mathbb{R}^{C \times C'}$ and $W_{TSA2} \in \mathbb{R}^{C' \times C}$ are learnable weight matrices and $C' = \max(8, 0.25C)$ is a reduced hidden dimension. The dimensionality reduction in the hidden layer serves two purposes: it decreases parameter overhead and encourages the model to learn compact saliency descriptors rather than overfitting to channel noise. Then the raw saliency score s_t is mapped into the interval (0,1) using a σ function:

$$\alpha_t = \sigma(s_t) \quad (13)$$

where α_t represented the learned temporal importance of the t-th time step. In this function, the Sigmoid function is used to process each time step independently, avoiding the occurrence of global competition and coupling phenomena [32]. The saliency-weighted feature is then reweighted element-wise which can be expressed in tensor form as:

$$X^{(s)} = X \odot \alpha_t \quad (14)$$

TSA module has achieved adaptive enhancement of information features and effectively suppressed the influence of signal noise on the inherent features of the signal.

4.2.2 Energy-aware Multi-Scale Aggregation mechanism

Although TSA highlights informative regions, effective modulation recognition also requires capturing signal patterns across different temporal receptive fields. Therefore, we further introduce an Energy-aware Multi-Scale Aggregation (EMA) mechanism.

For the input $X^{(s)} \in \mathbb{R}^{B \times T \times C}$ from TSA, it is first transposed to obtain a feature tensor of the form in $\mathbb{R}^{B \times T \times C}$. Then a multi-scale convolution using 1D convolution layers with kernel sizes $K_1 = 3$, $K_2 = 7$, $K_3 = 15$ is applied to $X^{(s)}$ for feature extraction to obtain F_{conv} , following the same operation process as in section 4.1.1.

Meanwhile, to further exploit the amplitude-frequency characteristics of the signal, an energy-guided gating mechanism is introduced. The channel-wise energy descriptor is defined as:

$$E_c = \frac{1}{T} \sum_{t=1}^T |X_{c,t}^{(s)}| \quad (15)$$

The use of the absolute value ensures that both positive and negative oscillatory components contribute to the energy estimation, which is consistent with the physical interpretation of signal magnitude in wireless communication systems. By averaging across the temporal dimension, E_c captures the overall activation strength of the c -th channel and reflects its contribution to representing modulation structures.

The energy descriptor is then transformed into a gating coefficient:

$$G_c = \sigma(E_c) \quad (16)$$

In this function, we employ the Sigmoid function to decouple channel-wise dependencies. In scenarios involving sophisticated modulation structures, critical information is frequently distributed across several sub-channels; the Sigmoid-based gating ensures that all segments exhibiting pronounced energy transients are emphasized simultaneously. This architectural flexibility prevents the inadvertent suppression of secondary yet vital signal components, ensuring a more holistic representation of the modulation manifold. Then the multi-scale feature F_{Conv} is reweighted channel-wise:

$$F_{c,t}^{(e)} = G_c F_{c,t} \quad (17)$$

Finally, a residual connection is introduced to preserve the original temporally refined representation and facilitate stable gradient propagation:

$$X^{(e)} = F^{(e)} + X^{(s)} \quad (18)$$

The residual connection acted as an additive gain mechanism stabilizes energy within the modulation feature, preventing attention gating from suppressing critical signal components which is a common issue in multiplicative scaling.

The ETSA module, integrating TSA and EMA as a unified gating mechanism within a progressive refinement framework, suppresses uninformative transients and dynamically re-calibrates the channel space via energy-aware modulation based on the

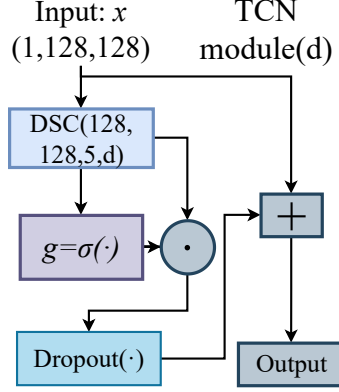


Fig. 5 TCN module

signal's physical power distribution, thereby ensuring the separability of the feature manifold under rapid channel fluctuations and impulsive interference.

4.3 Gated dilated Temporal Convolutional Network

To capture long-range temporal correlations beyond the local scope of ETSA, we employ the gated dilated Temporal Convolutional Networks (TCN) module characterized by gated dilated convolutions, as shown in Figure 5. The core advantage of this module lies in its ability to enlarge the effective temporal receptive field without introducing excessive computational overhead. And the integration of gated gating units (instead of standard ReLU) provides an additional layer of expressive nonlinearity, selectively amplifying modulation-relevant features while suppressing ambient channel artifacts. Furthermore, the residual-linked blocks serve as a structural backbone, safeguarding gradient integrity throughout the deep temporal modeling process[33].

For an input feature tensor X , each block computes:

$$F = f(X) \quad (19)$$

$$G = \sigma(g(X)) \quad (20)$$

$$H = F \odot G \quad (21)$$

where $f(\cdot)$ and $g(\cdot)$ denote separable DSC modules. The gated response is combined with the residual input:

$$X' = H + X \quad (22)$$

Three such blocks are stacked with dilation factors $d = 1, 2, 4$, enabling exponential growth of the effective receptive field.

4.4 Robust classifier

The final stage employs a classification head to map the learned features to modulation categories. Instead of flattening the temporal sequence, global average pooling layer is applied along the time dimension to generate a fixed-length representation. This aggregation summarizes the overall statistical characteristics of the signal while reducing sensitivity to temporal variations and short-term noise. Specially, the global temporal aggregation is first performed to derive the mean feature vector $\bar{X}$, which is formulated as :

$$\bar{X} = \frac{1}{T} \sum_{t=1}^T X_t \quad (23)$$

Following the aggregation of temporal information, the global feature vector $\bar{X}$ is fed into a dual-layer fully connected network structured to distill high-level semantic abstractions. The first transformation layer maps the consolidated features into a latent space of 128 dimensions utilizing GELU as the non-linear activation function. And to mitigate the risk of overfitting, a dropout mechanism is integrated immediately after the non-linear mapping, which is expressed as:

$$h = Dropout(GELU(W_{fc1} \cdot \bar{X} + b_{fc1})) \quad (24)$$

After the dropout mechanism, a layer normalization operation is employed which standardizes the activations across the feature dimension. To stabilize the distribution of the latent representation h before the final decision, the layer normalization is applied:

$$\hat{h} = LN(h) \quad (25)$$

Finally, the final linear layer serves as the decision boundary, projecting the normalized latent representation $\hat{h}$ onto the output space corresponding to the pre-defined modulation classes. The final prediction logic y generated by this output projection layer is expressed as:

$$y = W_{fc2}\hat{h} + b_{fc2} \quad (26)$$

5 Experimental Results

5.1 Dataset

The paper used the RML2016.10a, RML2018.01a, RML22 dataset to evaluate the modulation recognition task. The RML2016.10a and RML2018.01a datasets are generated using the GNU Radio signal generation toolchain and provided by DeepSig as part of the RadioML series [7, 34]. The RML22 dataset was proposed by Venkatesh Sathyanarayanan et al. and is a recent, more realistic RF signal dataset designed to better reflect practical wireless environments [35]. These datasets are public synthetic wireless signal datasets, have become standard evaluation resources in deep learning-based modulation classification research. The detailed characteristics of these datasets have been shown in Table 1.

To evaluate the proposed model, this paper divided datasets into training set, validation set and test set at a ratio of 70%, 20% and 10%, respectively. The network is trained for 100 epochs using the Adam optimizer with an initial learning rate of 0.001. Regarding the batch size configuration, it is set to 128 for the RML2016.10a and RML22 datasets and the batch size is adjusted to 512 for the larger-scale RML2018.01a dataset. The entire framework is implemented using PyTorch2.8.0 with CUDA12.6. All experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 4060 Laptop GPU.

Table 1 the detailed characteristics of RML2016.10a, RML2018.01a and RML22

Dataset	Sample Dimension	SNR range(dB)	Dataset Size	Modulation Schemes
RML2016.10a	128*2	-20:2:18	220,000	8PSK, BPSK, CPFSK, GFSK, PAM4, 16QAM, AM-DSB, AM-SSB, 64QAM, QPSK, WBFM (11 classes)
RML2018.01a	1024*2	-20:2:30	2,555,904	OOK, 4ASK, 8ASK, BPSK, QPSK, 8PSK, 16PSK, 32PSK, 16APSK, 32APSK, 64APSK, 128APSK, 16QAM, 32QAM, 64QAM, 128QAM, 256QAM, AM-SSB-WC, AM-SSB-SC, AM-DSB-WC, AM-DSB-SC, FM, GMSK, OQPSK (24 classes)
RML22	128*2	-20:2:20	420,000	8PSK, AM-DSB, BPSK, CPFSK, GFSK, PAM4, 16QAM, 64QAM, QPSK, WBFM (10 classes)

5.2 Comparison with existing models

In this section, this paper compared the proposed model EMST-Net with some representative deep learning-based AMC models from prior works, including CGDNet[36], MCCSAN[20], SCRNN[14], ResLSTM[37], CVCNN-LSTM[17], MoE-AMC[13], MCLDNN[16].

Table 2-4 compare the proposed model with other representative models on RML22, RML2016.10a and RML2018.01a datasets. On all datasets, the proposed model achieves the best overall performance in all metrics across the three datasets, demonstrating superior recognition performance. As illustrated in Table 2, our proposed model achieves an overall accuracy of 71.59%, a 0dB accuracy of 86.27%, a peak accuracy of 99.46%, a precision of 72.73%, and an F1-score of 0.7137. Compared to the competing models, the proposed approach yields substantial improvements: the average accuracy increased by 0.96%-8.38%, accuracy at 0dB SNR rose by 0.73%-19.74%, peak accuracy across varying SNR levels improved by 4.35%-7.23%, while precision and F1-score saw gains of 1.15%-7.24% and 2.66%-7.68%, respectively. What's more, the proposed model also performed significant improvement in all metrics in the Table

Table 2 Comparing the performance of models with RML22

Model	OA	PA	0dB Acc.	Precision	F1
CGDNet	0.6885	0.9835	0.8172	0.7056	0.6856
MCCSAN	<u>0.6878</u>	<u>0.9850</u>	<u>0.8192</u>	<u>0.7158</u>	<u>0.6871</u>
SCRNN	0.6743	0.9259	0.8176	0.6632	0.6590
ResLSTM	0.6814	0.9817	0.8106	0.6817	0.6787
CVCNN-LSTM	0.6464	0.9108	0.7904	0.6549	0.6369
MoE-AMC	0.6785	0.9765	0.8090	0.6884	0.6743
MCLDNN	0.6700	0.9389	0.8120	0.6705	0.6647
Our Proposed	0.7159	0.9946	0.8627	0.7273	0.7137

Notes: OA represents the overall classification accuracy, PA represents the peak classification accuracy, 0dB Acc. Represents the classification accuracy with the 0dB SNR.

Table 3 Comparing the performance of models with RML2016.10a

Model	OA	PA	0dB Acc.	Precision	F1
CGDNet	<u>0.6150</u>	<u>0.9185</u>	<u>0.8780</u>	<u>0.7312</u>	<u>0.6351</u>
MCCSAN	0.5974	0.8888	0.8482	0.7009	0.6167
SCRNN	0.5733	0.8488	0.8188	0.6640	0.5720
ResLSTM	0.6035	0.8966	0.8608	0.6866	0.6176
CVCNN-LSTM	0.5712	0.8398	0.8252	0.6763	0.5763
MoE-AMC	0.4828	0.7284	0.6621	0.5318	0.4632
MCLDNN	0.5981	0.8959	0.8627	0.7061	0.7049
Our Proposed	0.6279	0.9258	0.8965	0.7497	0.6477

Table 4 Comparing the performance of models with RML2018.01a

Model	OA	PA	0dB Acc.	Precision	F1
CGDNet	0.5555	0.8743	0.5112	0.5994	0.5630
MCCSAN	<u>0.6052</u>	<u>0.9544</u>	0.5388	0.6356	0.6069
SCRNN	0.4838	0.7871	0.4416	0.5966	0.4930
ResLSTM	0.5941	0.9449	<u>0.5443</u>	0.6187	0.5963
CVCNN-LSTM	0.4741	0.7190	0.5073	0.4949	0.4575
MoE-AMC	0.5074	0.8162	0.4804	0.5978	0.5219
MCLDNN	0.6028	0.9443	0.5353	<u>0.6477</u>	<u>0.6080</u>
Our Proposed	0.6174	0.9738	0.5605	0.6963	0.6314

3 and 4. The above data results and analysis have proved that the proposed model EMST-Net is effective in improving the modulation recognition performance.

Beyond the classification performance, we also need to analyze the computational complexity and resource requirements of the models to assess its feasibility for deployment on the edge. Table 5 shows the computational complexity and storage overhead of the proposed EMST-Net and other comparative models. EMST-Net operates with

Table 5 Comparing the computational complexity and storage overhead of models

Model	FLOPs (M)	Param (K)	Storage (MB)
CGDNet	16.689	427.979	1.639
MCCSAN	54.347	455.755	1.739
SCRNN	16.098	<u>232.651</u>	0.889
ResLSTM	<u>15.533</u>	361.675	1.383
CVCNN-LSTM	45.932	405.847	1.550
MoE-AMC	15.133	128.376	0.620
MCLDNN	22.921	370.787	1.417
Our proposed	24.245	260.078	<u>0.811</u>

Notes: FLOP represents floating-point operations per second and Param represents parameters of the model.

24.245M FLOPs, 260.078K parameters, and a storage footprint of 0.811 MB. This resource consumption meets practical deployment requirements.

In terms of computational complexity (FLOPs), MCCSAN and CVCNN-LSTM exhibit resource consumption and require 54.347M and 45.923M FLOPs. Both are computation-intensive models, suffering from significant computational bottlenecks, and limit their applicability in resource-contained scenarios. EMST-Net with 24.245M FLOPs is 55.4% and 47.2% lower than MCCSAN and CVCNN-LSTM. We acknowledge a marginal inference premium relative to ultra-lightweight baselines like MoE-AMC and SCRNN. However, this modest computational increment is a deliberate design choice, strictly allocated to power the multi-scale spatial encoding and energy-guided temporal modeling. Ultimately, EMST-Net introduces no excessive structural redundancy, navigating an optimal Pareto front between discriminative abstraction and resource efficiency for real-world scenarios.

Figure 6-8 illustrate the modulation classification accuracy, precision and F1 of the models under different SNR conditions. As shown in the figures, EMST-Net outperforms the comparative models across all SNR levels. These result demonstrates that the proposed model not only exhibits strong feature extraction capability under high-quality signal conditions but also maintains remarkable robustness in extremely low-SNR environments. This outstanding performance can be attributed to the synergistic interaction among the model components, particularly the EMST-Net module, which enhances intrinsic signal characteristics while effectively suppressing noise interference.

5.3 Comparison on Attention Mechanisms

To verify the effectiveness of ETSA module, we replaced this component with other attention mechanism for comparison experiments, including CBAM [38], Hyena [39] and Multi-head [40]. The experimental results are shown in Table 6. As shown in Table 6, ETSA demonstrated the best performance in terms of ETSA, precision and F1 score. Compared with other attention mechanism, the OA increased by 1.09-1.57%,

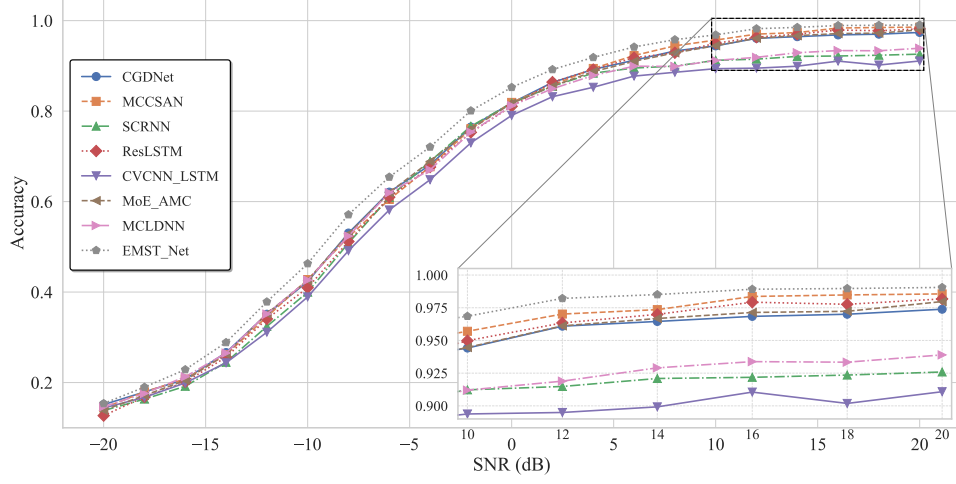


Fig. 6 Classification accuracy comparison between the proposed model and compared models on the RML22 dataset.

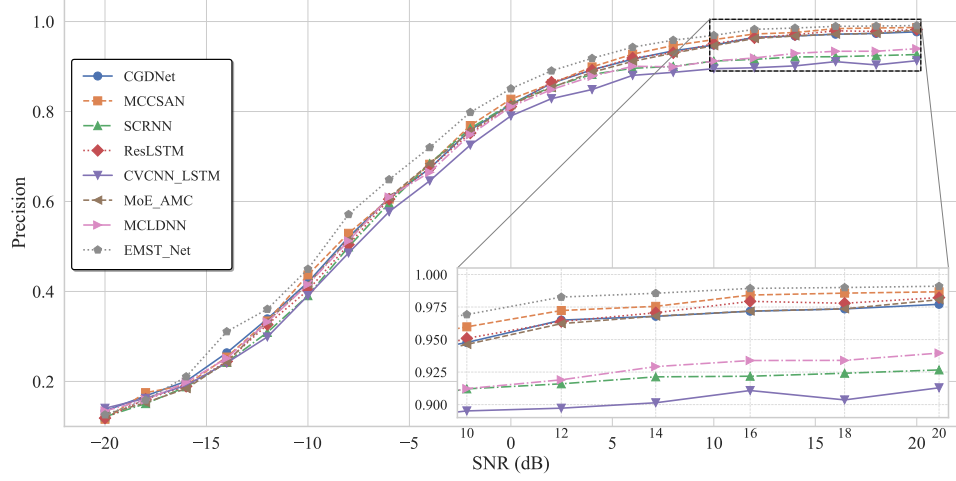


Fig. 7 Classification precision comparison between the proposed model and compared models on the RML22 dataset.

the precision rate increased by 1.03%-1.28%, and the F1 score increased by 1.25%-1.68%. Additionally, our proposed model also had the lowest computational resource requirements, with FLOPs being only 30.622M and the parameter quantity being only 260.078K.

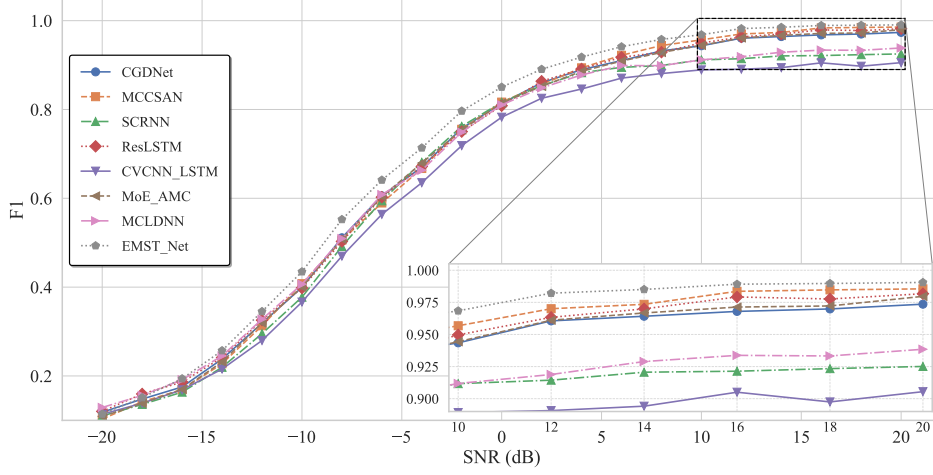


Fig. 8 F1 score comparison between the proposed model and compared models on the RML22 dataset.

Table 6 Comparison on Attention Mechanisms on the RML22 dataset.

Model	OA	Precision	F1	FLOPs	Param(K)
CBAM	0.7002	<u>0.7146</u>	0.6969	<u>31.781</u>	<u>272.827</u>
Hyena	<u>0.7050</u>	0.7145	<u>0.7012</u>	44.462	368.813
Multi-Head	<u>0.7020</u>	0.7171	0.6989	40.150	334.381
ETSA	0.7159	0.7273	0.7137	30.622	260.078

5.4 Ablation experiment

In the proposed model, it is combined with two significant modules: Multi-Stream Fusion Convolution Module (MSFC) and Energy-Temporal Saliency Attention Module (ETSA). MSFC module mainly adopts multi-scale convolution (MC) module and DFR. ETSA module adopts TSA and EMA. In this section, we conducted ablation experiments to verify whether these modules could actually play a significant role in enhancing the performance of the model.

As shown in Figure 7, introducing MC (M2) leads to a noticeable improvement over the baseline, demonstrating the effectiveness of multi-scale spatial feature extraction. Similarly, DFR (M3) enhances Precision, while TSA (M4) and EMA (M5) each provide consistent gains, confirming the benefits of temporal saliency modeling and energy-aware enhancement. When the model combines MC and DFR together (M6), or combines TSA and EMA together (M7), the performance further improves, indicating that they have a complementary effect rather than being redundant. Finally, the complete model achieves the best results, reflecting the synergistic gains accumulated from the integration of all components, and validating the effectiveness of the

overall architecture. This demonstrates that each module complements one another and jointly enhances model performance.

Table 7 Ablation Study on the Effectiveness of Core Modules on the RML22 dataset.

Model	MC	DFR	TSA	EMA	Accuracy	Precision	F1
M1	-	-	-	-	0.6942	0.7155	0.6911
M2	✓	-	-	-	0.7054	0.7171	0.7029
M3	-	✓	-	-	0.7040	<u>0.7250</u>	0.7029
M4	-	-	✓	-	0.6986	0.7093	0.6947
M5	-	-	-	✓	0.7018	0.7199	0.6993
M6	✓	✓	-	-	<u>0.7088</u>	0.7245	<u>0.7065</u>
M7	-	-	✓	✓	0.7068	0.7220	0.7042
Our proposed	✓	✓	✓	✓	0.7159	0.7273	0.7137

5.5 Classification performance of the model for different modulation schemes

Figure 9 shows the per-modulation scheme accuracy of EMST-Net on the RML22 dataset. Overall, the accuracy of all modulation schemes increases monotonically with the improvement of SNR. When the SNR is less than -10dB, the accuracy for most modulations remains below 0.5, indicating the noise significantly interferes with signal characteristics. In contrast, when the SNR is more than 10dB, the accuracy for all schemes approach 1, demonstrating that the model can achieve stable classification with weak interference channel.

Specially, the performance of different modulation schemes have different manifestations. Modulation schemes such as AM-SSB, BPSK, CPFSK, GFSK, and PAM4 exhibit high robustness, achieving relatively high accuracy (over 0.5) even at extremely low SNRs (e.g., -20 dB) and showing the steepest rise in accuracy, approaching nearly perfect recognition at SNRs close to 0 dB. This indicates that the signal features of these modulations are less sensitive to noise and can be easily captured and distinguished by the model. Modulation methods such as AM-DSB and WBFM perform less well than the aforementioned ones. They can maintain stable performance within the medium signal-to-noise ratio range (-10 dB to 0 dB), and can achieve an accuracy of above 0.9 only when the signal-to-noise ratio is greater than 5 dB.

Compared to the aforementioned modulation methods, higher-order modulation methods such as 8PSK, QAM16, QAM64 and QPSK are more sensitive to noise. Their accuracy rates are usually lower than 0.5 when the signal-to-noise ratio is below 0 decibels, and only start to significantly improve when the signal-to-noise ratio exceeds 5 decibels. Moreover, QAM64 performs the worst. Even at a signal-to-noise ratio of 0 decibels, its accuracy rate is still lower than 0.5.

Further, we plotted a heatmap of the classification accuracy for each modulation scheme at different SNRs to more intuitively present the performance variation of

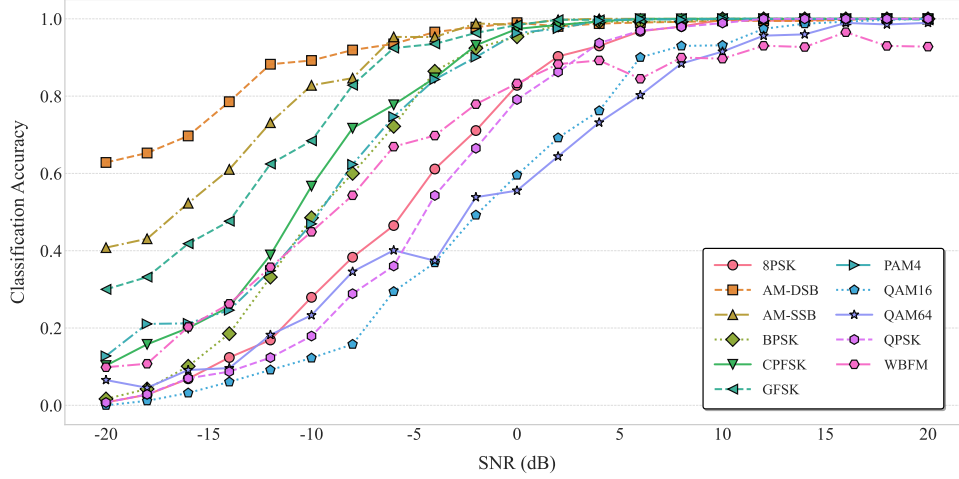


Fig. 9 Classification accuracy analysis of different modulation schemes with EMTE-Net Under Different SNR Conditions

different modulations with respect to SNR, as shown in Figure 10. Figure 10 provides a clear quantitative visualization of recognition accuracy across SNRs. At low SNRs, high-order modulations including 8PSK, QAM16, QAM64, and QPSK show extremely low accuracy, while AM-SSB and AM-DSB maintain relatively high values, demonstrating stronger noise immunity. With the improvement of SNR, the accuracy of BPSK, CPFSK, GFSK, and PAM4 improves rapidly, whereas high-order modulations recover much more slowly. At high SNRs, most modulation schemes achieve nearly perfect accuracy, except that QAM64 and WBFM remain slightly lower. This result coincides exactly with Figure 9.

5.6 Confusion matrix and t-SNE visualization analysis

To provide a comprehensive analysis of class-wise recognition performance, Figure 11 presents the normalized confusion matrices under different SNR conditions. In Figure 11, the figure is depicted from four different perspectives, including overall SNR, low SNR (-20 0 dB), medium SNR (0 10 dB), and high SNR (10 20 dB).

From the overall SNR matrix, a strong diagonal dominance can be clearly observed, indicating that the proposed model learns highly discriminative representations from IQ signals. Constant-envelope modulations, including BPSK, CPFSK, and GFSK, achieve near-perfect recognition accuracy, indicating that the network effectively captures their unique temporal characteristics. Analog modulations (e.g., AM-DSB and AM-SSB) are also identified with high accuracy, reflecting reliable extraction of envelope-related features.

However, under low SNR conditions, performance degradation is observable due to severe noise distortion; however, the diagonal structure remains largely preserved,

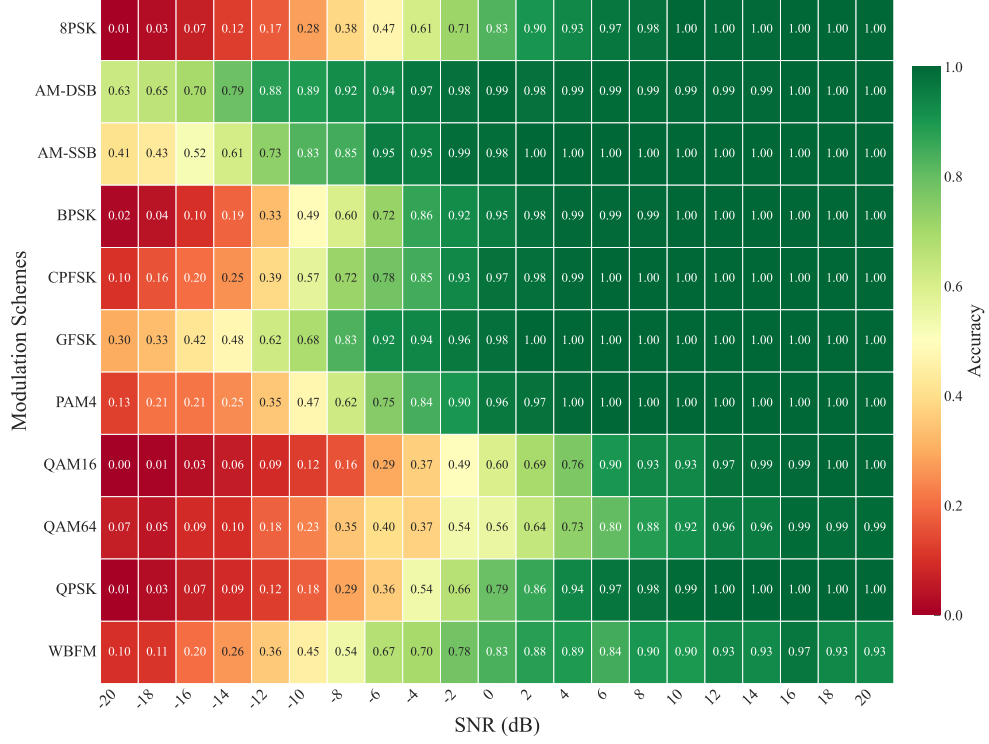


Fig. 10 The heatmap of the classification accuracy for each modulation scheme at different SNRs

reflecting strong robustness. The primary misclassifications occur between modulation schemes with inherently similar constellation geometries, particularly QAM16 and QAM64, as well as among phase-shift keying variants. What's more, when the SNR increases to the medium range, off-diagonal errors are significantly reduced and the diagonal patterns become sharper, indicating improved separability in both amplitude and phase domains. At high SNR levels, the confusion matrix approaches an identity matrix, demonstrating near-complete class separability. Only minimal confusion remains among high-order modulation formats with closely related statistical properties.

Then, we performed t-SNE visualization under different SNR conditions, as shown in Figure 12. From the overall SNR distribution, most modulation types form relatively distinguishable clusters, although partial overlaps remain among modulation schemes with similar structural properties, consistent with the minor confusions observed in the corresponding confusion matrix. Under low SNR conditions, feature distributions become dispersed and inter-class boundaries blur due to severe noise perturbation, especially high-order QAM formats and phase-based modulations. As the SNR increases, intra-class compactness improves and inter-class separation becomes

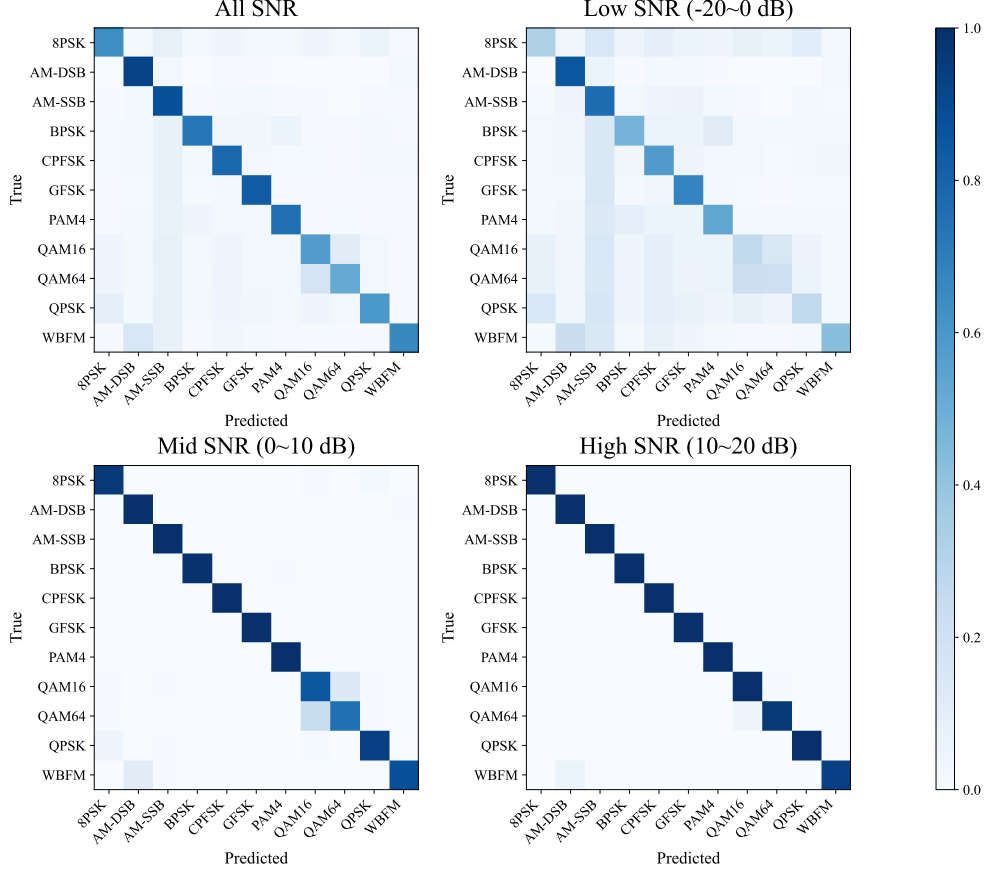


Fig. 11 Confusion matrices of the proposed model under different SNR conditions (All SNR, -20 0 dB, 0 10 dB, and 10 20 dB) on the RML22 dataset.

more pronounced, indicating that the model learns progressively stable modulation-specific representations. Finally, at high SNR levels, clusters are well compacted with clear margins between categories.

6 Discussion

This paper proposes an Energy-Guided Multi-Scale Temporal Network (EMST-Net) for AMC. The model enhances feature expressiveness and separability through three core components: a Multi-stream Fusion Convolution module (MSFC) that decouples joint IQ signals from their independent components, an Energy-Guided Temporal Saliency Attention module (ETSA) that dynamically enhances informative signal segments while suppressing noise, and a Gated Dilated Temporal Convolutional Network (TCN) that captures long-range temporal dependencies. Experimental evaluations on public datasets demonstrate that EMST-Net consistently outperforms existing

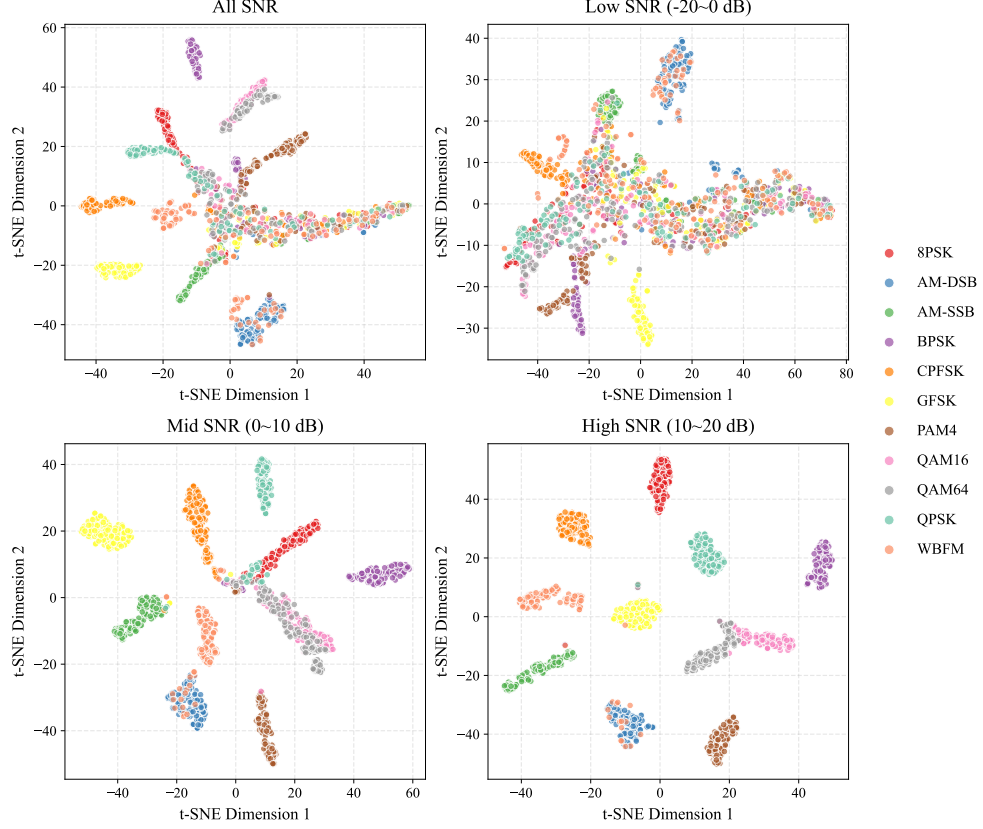


Fig. 12 t-SNE visualization of proposed model-extracted features under different SNR conditions (All SNR, -20 0 dB, 0 10 dB, and 10 20 dB) on the RML22 dataset.

DL models across all SNR while maintaining a lightweight architecture suitable for edge deployment. Then ablation experiments confirm the contribution of each module. Finally, confusion matrices verify the model’s classification accuracy, and t-SNE demonstrates the separability of learned features under varying SNR conditions.

Despite the superior performance of EMST-Net in multiple aspects, certain limitations remain. The current study is primarily based on synthetic datasets for training and evaluation. Although these datasets are widely recognized in modulation classification research, factors such as multipath effects, interference signals, and hardware non-idealities in real-world wireless channels are not fully modeled, and the generalization capability of the model in real environments requires further validation. Additionally, the recognition capability of the current model for higher-order modulation formats under low SNR remains limited. Future work could consider incorporating richer time-frequency features or leveraging generative adversarial samples to enhance robustness.

Declarations

- Funding
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.
- Competing interests
The authors declare that they have no conflict of interest.
- Ethics approval and consent to participate
This study does not involve human participants or animals.
- Consent for publication
All authors have approved the manuscript and agree with its submission for publication.
- Data availability
The RML2018.01A dataset and RML2016.10a dataset used in this study were released by DeepSig and are publicly available at <https://www.deepsig.ai/datasets>. The RML22 dataset was released by venkateshsathya and is publicly available at <https://github.com/venkateshsathya/RML22>.
- Materials availability
Not applicable.
- Code availability
The code used in this study is available from the corresponding author upon reasonable request.
- Author contribution
Jiahua Zhou: Conceptualization, methodology, experiments, and writing—original draft preparation. The author reviewed and approved the final manuscript.

References

- [1] Zheng, Q., Tian, X., Yu, L., Elhanashi, A., Saponara, S.: Recent advances in automatic modulation classification technology: Methods, results, and prospects. *International Journal of Intelligent Systems* **2025**(1), 4067323 (2025)
- [2] Yang, W., Sun, X., Kang, K., Dai, J., Zhao, P.: Wavelet convolutional neural network and dual-channel feature extraction for automatic modulation classification. *Journal of King Saud University Computer and Information Sciences* (2025)
- [3] Gu, S., Su, J., Liu, C., Cheng, X.: A lightweight multi-channel feature fusion approach for automatic modulation recognition. *Signal, Image and Video Processing* **19**(13), 1154 (2025)
- [4] Oikonomou, T.K., Evgenidis, N.G., Nixarlidis, D.G., Tyrovolas, D., Tegos, S.A., Diamantoulakis, P.D., Sarigiannidis, P.G., Karagiannidis, G.K.: Cnn-based automatic modulation classification under phase imperfections. *IEEE Wireless Communications Letters* **13**(5), 1508–1512 (2024)

- [5] Zhang, F., Luo, C., Xu, J., Luo, Y., Zheng, F.-C.: Deep learning based automatic modulation recognition: Models, datasets, and challenges. *Digital Signal Processing* **129**, 103650 (2022)
- [6] Hazza, A., Shoaib, M., Alshebeili, S.A., Fahad, A.: An overview of feature-based methods for digital modulation classification. In: 2013 1st International Conference on Communications, Signal Processing, and Their Applications (ICCSPA), pp. 1–6 (2013). IEEE
- [7] O’Shea, T.J., Corgan, J., Clancy, T.C.: Convolutional radio modulation recognition networks. In: International Conference on Engineering Applications of Neural Networks, pp. 213–226 (2016). Springer
- [8] Hong, D., Zhang, Z., Xu, X.: Automatic modulation classification using recurrent neural networks. In: 2017 3rd IEEE International Conference on Computer and Communications (ICCC), pp. 695–700 (2017). IEEE
- [9] Rajendran, S., Meert, W., Giustiniano, D., Lenders, V., Pollin, S.: Deep learning models for wireless signal classification with distributed low-cost spectrum sensors. *IEEE Transactions on Cognitive Communications and Networking* **4**(3), 433–445 (2018)
- [10] Xie, T., Han, X.: Modulation pattern recognition method of wireless communication automatic system based on iabln algorithm in intelligent system. *PLoS One* **20**(1), 0317355 (2025)
- [11] West, N.E., O’shea, T.: Deep architectures for modulation recognition. In: 2017 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), pp. 1–6 (2017). IEEE
- [12] Zou, B., Zeng, X., Wang, F.: Research on modulation signal recognition based on cldnn network. *Electronics* **11**(9), 1379 (2022)
- [13] Gao, J., Cao, Q., Chen, Y.: Moe-amc: Enhancing automatic modulation classification performance using mixture-of-experts. *arXiv preprint arXiv:2312.02298* (2023)
- [14] Liao, K., Zhao, Y., Gu, J., Zhang, Y., Zhong, Y.: Sequential convolutional recurrent neural networks for fast automatic modulation classification. *IEEE Access* **9**, 27182–27188 (2021)
- [15] Singh, R.P., Singh, L.D.: Dyhand: dynamic hand gesture recognition using bilstm and soft attention methods. *The Visual Computer* **41**(1), 41–51 (2025)
- [16] Xu, J., Luo, C., Parr, G., Luo, Y.: A spatiotemporal multi-channel learning framework for automatic modulation recognition. *IEEE Wireless Communications Letters* **9**(10), 1629–1632 (2020)

- [17] Cheng, R., Chen, Q., Huang, M.: Automatic modulation recognition using deep cvcnn-lstm architecture. *Alexandria Engineering Journal* **104**, 162–170 (2024)
- [18] Lin, C., Zou, C., Xu, H.: Scnet: A dual-branch network for strong noisy image denoising based on swin transformer and convnext. *Computer Animation and Virtual Worlds* **36**(3), 70030 (2025)
- [19] Senhua, X., Liqing, G., Liang, W., Wei, F.: Multi-scale context-aware network for continuous sign language recognition. *Virtual Reality & Intelligent Hardware* **6**(4), 323–337 (2024)
- [20] Xu, S., Zhang, D., Lu, Y., Xing, Z., Ma, W.: Mccsan: Automatic modulation classification via multiscale complex convolution and spatiotemporal attention network. *Electronics* **14**(16), 3192 (2025)
- [21] Zhang, S., Song, Y., Wang, S.: Automatic modulation classification via recurrent self-attention with weight non-negative constraint. *IET Communications* **19**(1), 70025 (2025)
- [22] Wu, T., Zheng, W., Liu, L.: Automatic modulation recognition using vision transformers with cross-layer feature fusion and kolmogorov–arnold representation. *Wireless Networks* **31**(6), 3981–3992 (2025)
- [23] Pei, Y., Liu, L., Jiang, R., Zhang, Y., Yu, P., Lin, L., Guo, Y.: Energy-guided test-time adaptation for data shifts in multi-modal perception. *The Visual Computer* **41**(9), 6533–6546 (2025)
- [24] Wen, Y., Luo, B., Shi, W., Ji, J., Cao, W., Yang, X., Sheng, B.: Sat-net: structure-aware transformer-based attention fusion network for low-quality retinal fundus images enhancement. *IEEE Transactions on Multimedia* (2025)
- [25] Wen, Y., Wang, X.-N., Tang, J., Li, P., Zhu, L., Qin, J., Yang, X., Sheng, B.: Ccm-net: Contrastive and consistent multi-task network for artifact segmentation and quality classification of octa images. *ACM Transactions on Multimedia Computing, Communications and Applications* **21**(12), 1–21 (2025)
- [26] Kong, W., Jiao, X., Xu, Y., Yang, Q.: An effective masked transformer model for automatic modulation recognition. *IEEE Transactions on Cognitive Communications and Networking* (2025)
- [27] Li, Y., Shi, X., Tan, H., Zhang, Z., Yang, X., Zhou, F.: Multi-representation domain attentive contrastive learning based unsupervised automatic modulation recognition. *Nature Communications* **16**(1), 5951 (2025)
- [28] Zhang, H., Zhou, F., Wu, Q., Wu, W., Hu, R.Q.: A novel automatic modulation classification scheme based on multi-scale networks. *IEEE Transactions on Cognitive Communications and Networking* **8**(1), 97–110 (2021)

- [29] Uddin, M.R., Ullah, M.S., Islam, M.S.: Radionet: A multi-stream deep attentive network for low-snr automatic modulation classification
- [30] Zhang, Q., Dong, X., Xiong, J., Zhu, H., Wang, Q., Wang, J., Dai, W.: Mcam-net: multi-scale convolutional attention for enhanced industrial surface defect detection. *The Visual Computer* **42**(4), 182 (2026)
- [31] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017)
- [32] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018)
- [33] Hao, H., Wang, Y., Xue, S., Xia, Y., Zhao, J., Shen, F.: Temporal convolutional attention-based network for sequence modeling. *arXiv preprint arXiv:2002.12530* (2020)
- [34] O’Shea, T.J., Roy, T., Clancy, T.C.: Over-the-air deep learning based radio signal classification. *IEEE Journal of Selected Topics in Signal Processing* **12**(1), 168–179 (2018)
- [35] Sathyanarayanan, V., Gerstoft, P., El Gamal, A.: Rml22: Realistic dataset generation for wireless modulation classification. *IEEE Transactions on Wireless Communications* **22**(11), 7663–7675 (2023)
- [36] Njoku, J.N., Morocho-Cayamcela, M.E., Lim, W.: Cgdnet: Efficient hybrid deep learning model for robust automatic modulation recognition. *IEEE Networking Letters* **3**(2), 47–51 (2021)
- [37] Elzagheer, M.M., Ramzy, S.M.: A hybrid model for automatic modulation classification based on residual neural networks and long short term memory. *Alexandria Engineering Journal* **67**, 117–128 (2023)
- [38] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19 (2018)
- [39] Poli, M., Massaroli, S., Nguyen, E., Fu, D.Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., Ré, C.: Hyena hierarchy: Towards larger convolutional language models. In: *International Conference on Machine Learning*, pp. 28043–28078 (2023). PMLR
- [40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)